PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-163248
(43)Date of publication of application : 07.06.2002

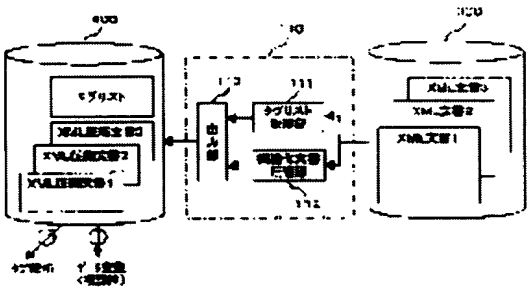
(51)Int.Cl. G06F 17/21
G06F 12/00
G06F 17/30

(21)Application number : 2000-357568
(22)Date of filing : 24.11.2000

(71)Applicant : FUJITSU LTD
(72)Inventor : SATO NOBUKO

(54) STRUCTURED DOCUMENT COMPRESSOR, STRUCTURED DOCUMENT RESTORING DEVICE AND STRUCTURED DOCUMENT PROCESSING SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To reduce the data amount of a structured document by compressing the structured document while keeping the visibility, flexibility and expandability of the data structure which are advantages of the structure document, and to reduce load of tag analysis by dispensing with wasteful tag analysis in han dling a number of structured documents having the same data structure.
SOLUTION: This structured document processing system is provided with a tag list generating part 111 for generating one tag list by extracting tags in the structured documents in the appearing order to make a list, a structured document compressing part 112 for generating a compressed document by replacing a tag in each structure document with a designated partition code, and an output part 113 for making the above one tag list generated by the tag list generating part 111 correspond to a plurality of compressed documents generated by the structured document compressing part 112 concerning each of the plurality of structured documents and outputting the same as the compression result.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－163248
（P2002－163248A）
(43)公開日　平成14年６月７日(2002.6.7)

(51)Int.Cl.⁷	識別記号	FI	テーマコード（参考）
G06F 17/21	501	G06F 17/21	501T　5B009
	570		570G　5B075
			570D　5B082
12/00	511	12/00	511A
17/30	230	17/30	230A

審査請求　未請求　請求項の数５　ＯＬ　（全 25 頁）

(21)出願番号　特願2000－357568(P2000－357568)

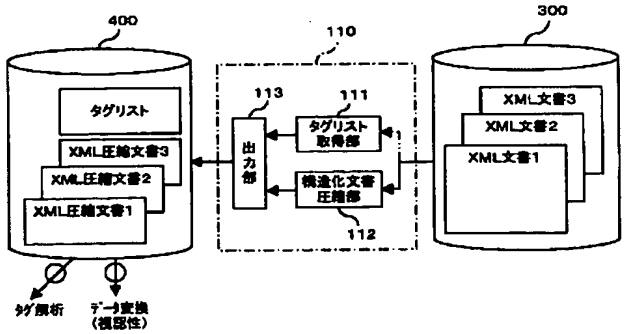
(22)出願日　平成12年11月24日(2000.11.24)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(72)発明者　佐藤　宣子
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(74)代理人　100092978
弁理士　真田　有

Ｆターム(参考)　5B009 NA05 SA08
5B075 NR03 NR16
5B082 GA01

(54)【発明の名称】　構造化文書圧縮装置および構造化文書復元装置並びに構造化文書処理システム

(57)【要約】
【課題】構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま構造化文書を圧縮できるようにして、構造化文書のデータ量の削減をはかるとともに、同一のデータ構造を有する多数の構造化文書を取り扱う際に無駄なタグ解析を行なう必要を一切なくして、タグ解析の負荷の低減をはかる。
【解決手段】構造化文書におけるタグを出現順序に従って抽出してリスト化した一つのタグリストを生成するタグリスト生成部１１１と、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部１１２と、タグリスト生成部１１１により生成された前記一つのタグリストと複数の構造化文書のそれぞれについて構造化文書圧縮部１１２により生成された複数の圧縮文書とを対応させ圧縮結果として出力する出力部１１３とをそなえる。



1

【特許請求の範囲】

【請求項１】　同一のデータ構造を有する複数の構造化文書を圧縮する装置であって、
該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、
各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、
該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部とをそなえたことを特徴とする、構造化文書圧縮装置。

【請求項２】　構造化文書を圧縮する装置であって、
該構造化文書中のタグを検出するタグ検出部と、
該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえたことを特徴とする、構造化文書圧縮装置。

【請求項３】　同一のデータ構造を有する複数の構造化文書中のタグを所定の区切りコードに置き換えることにより生成された複数の圧縮文書を、該複数の構造化文書におけるタグを出現順序に従ってリスト化したタグリストに基づいて復元する装置であって、
該タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製する複製部と、
該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえたことを特徴とする、構造化文書復元装置。

【請求項４】　構造化文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する装置であって、
該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、
該圧縮文書中の前記所定の区切りコードを検出する区切りコード検出部と、
該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえたことを特徴とする、構造化文書復元装置。

【請求項５】　同一のデータ構造を有する複数の構造化文書に対する処理を行なうべく、該複数の構造化文書を圧縮する構造化文書圧縮装置と、該構造化文書圧縮装置による圧縮データを該複数の構造化文書に復元する構造化文書復元装置とを含んで構成される構造化文書処理システムにおいて、

2

該構造化文書圧縮装置が、
該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、
各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、
該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部とをそなえて構成されるとともに、
該構造化文書復元装置が、
該複数の圧縮文書の復元結果を格納するメモリと、
該タグリストに対応するデータ構造を該メモリ上に複製データ構造として展開・複製する複製部と、
該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえて構成されたことを特徴とする、構造化文書処理システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ＳＧＭＬ(Standard Generalized Markup Language)やＸＭＬ(eXtensible Markup Language)等の構造化文書のための圧縮・復元に係る技術に関し、特に、同一のデータ構造を有する多数の構造化文書、例えば同一フォーマットの多数の伝票類を取り扱う際に用いて好適の、構造化文書圧縮装置および構造化文書復元装置並びに構造化文書処理システムに関する。

【０００２】

【従来の技術】近年、文字コード，画像データ等の様々な種類のデータがコンピュータで扱われるようになるに伴い、取り扱われるデータ量も増大している。そのような大量のデータをそのまま取り扱うと、そのデータを記憶するために多大な記憶容量が必要になり、そのデータの伝送に多大な時間を要することになるが、そのデータ中の冗長な部分を省いて圧縮すれば、記憶容量を減らしたり、遠隔地への伝送を高速化したりすることができる。様々なデータを一つの方式で圧縮することができる方法として、例えばユニバーサル符号化が提案されている。

【０００３】一般的なユニバーサルデータ圧縮方式としては、データ系列の類似性を利用した辞書型符号化方式と、データ列の出現頻度を利用した確率統計型符号化方式とがある（例えば、ＣＱ出版社刊：植松友彦著“文書データ圧縮アルゴリズム入門”参照）。いずれの方式においても、生成される圧縮データは、バイナリコード（例えば、８ビットコードで見ると、０ｘ００～０ｘＦＦの全てのコードを使用する）になっている。例えば図

１９は一般的なユニバーサルデータ圧縮について説明するための図であるが、この図１９に示すように、ユニバーサルデータ圧縮により、ＸＭＬ文書１，２，３はそれぞれバイナリデータ１，２，３に圧縮・変換されることになる。

【０００４】一方、最近では計算機で取り扱うデータの形式を統一する動きがある。即ち、これまで、計算機やアプリケーションによってバラバラであったデータ形式を、異なる計算機やアプリケーションでも使用できるようにするものである。このようなデータ形式を統一するための規格として、ＸＭＬが、１９９８年２月にＷ３Ｃ(World Wide Web Consortium)によって正式に勧告されている。このＸＭＬは、同様の規格ＳＧＭＬのサブセットになっており、文書自身の中にタグを埋め込む形で、その文書のデータ構造が記述される。ＸＭＬやＳＧＭＬにより記述された文書は、一般に構造化文書と呼ばれる。

【０００５】このような構造化文書は、データ構造をタグとして文書に埋め込んだ構成を採用しているので、データ構造について高い柔軟性／拡張性を有するという利点が得られる。また、人が見て意味のあるテキストによってタグが記述されているので、ＸＭＬ文書のデータ構造の視認性が高い。従って、データ交換を行ないやすくなり、複数種類の記述方法を緩やかに統合することができるという利点も得られる。このことは、構造化文書を成すデータが、アプリケーションに依存しないことを意味する。

【０００６】以下では、ＸＭＬ規格に従って、「＜」と「＞」とで囲まれた文字列（要素名を表す文字列）をタグ、「＜文字列＞」を開始タグ、「＜／文字列＞」を終了タグ、これらの開始タグと終了タグとの間にはさまれた領域に記述された文字列を要素（もしくは要素内容）と呼ぶ。

【０００７】ＸＭＬ規格に従って記述されたＸＭＬ文書の利用は、ウェブ(Web)やデータベースの分野を中心に増えつつあり、特に、ＸＭＬ文書は、ＥＤＩ(Electric DataInterchange)，ＥＣ(Electric Commerce)，携帯電話サービス，デジタルテレビ向けサービス，Ｗｅｂサービスなどで広く利用されつつある。図２０は、上述のようなＸＭＬ文書を取り扱う一般的なシステムの構成例を示すブロック図である。この図２０に示すシステムでは、ハードディスク（データベース）１０，読出部２０，メモリ展開部３０およびメモリ４０がそなえられている。

【０００８】ハードディスク（データベース）１０は、ＸＭＬ文書（図２０では３つのＸＭＬ文書１，２，３）を格納・保持するものであり、読出部２０は、ハードディスク１０のＸＭＬ文書をメモリ展開部３０へ読み出すものである。メモリ展開部３０は、読出部２０から入力されたＸＭＬ文書を、メモリ４０上に展開するためのもので、解析部３１，生成部３２および格納部３３を有して構成されている。

【０００９】ここで、解析部３１は、メモリ４０上に展開すべき各ＸＭＬ文書におけるタグを解析することにより、各ＸＭＬ文書のデータ構造（木構造）を解析するものであり、生成部３２は、解析部３１によるデータ構造解析結果に従って、各ＸＭＬ文書に応じた文書データを生成するものであり、格納部３３は、生成部３２により生成された文書データをメモリ４０上に展開し格納するものである。

【００１０】上述の構成により、図２０に示すシステムでは、ハードディスク１０に格納されているＸＭＬ文書が、読出部２０により読み出されてメモリ展開部３０に入力されると、解析部３１により、そのＸＭＬ文書のデータ構造（木構造）が解析される。そして、生成部３２において、解析部３１によるデータ構造解析結果に従い、そのＸＭＬ文書に応じた文書データが生成され、生成された文書データが、格納部３３によりメモリ４０上に展開されて格納される。

【００１１】

【発明が解決しようとする課題】ところで、ＸＭＬ文書（構造化文書）は、データ構造について高い柔軟性／拡張性を有するとともに高い視認性を有するが、人が見て意味を理解できるようにタグを記述するため、冗長な記述となり、そのＸＭＬ文書のデータ量は大幅に増大する。従って、ＸＭＬ文書を圧縮して、データ量を削減することが望まれている。

【００１２】しかしながら、ユニバーサルデータ圧縮を用いると、例えば図１９に示すごとく、圧縮データが全てバイナリデータになるため、ＸＭＬ文書の利点の一つであるデータの視認性が全くなくなってしまい、圧縮データを元の状態に復元しなければ、その圧縮データの内容を一切把握することができない。このため、圧縮データの復元アルゴリズムが分からなければ、データ交換もタグ解析も行なうことができない。

【００１３】一方、図２０を参照しながら説明したごとくＸＭＬ文書をメモリ４０上に展開する際、ＸＭＬ文書（構造化文書）ではデータ構造がタグとしてＸＭＬ文書中に埋め込まれているため、解析部３１によるタグ解析処理（データ構造解析処理）の負荷が高くなる。特に、同一のデータ構造を有する多数のＸＭＬ文書（例えば発注伝票）をメモリ４０上に展開する場合、全てのＸＭＬ文書が同一のデータ構造を有しているにもかかわらず、ＸＭＬ文書一つ一つについてデータ構造解析処理を行なわなければならず、無駄なタグ解析処理の量が増大し、その処理による負荷が極めて高くなるという課題があった。

【００１４】本発明は、このような課題に鑑み創案されたもので、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を

圧縮できるようにして、構造化文書のデータ量の削減をはかるとともに、同一のデータ構造を有する多数の構造化文書を取り扱う際に無駄なタグ解析を行なう必要を一切なくして、タグ解析の負荷の低減をはかった、構造化文書圧縮装置および構造化文書復元装置並びに構造化文書処理システムを提供することを目的とする。

【００１５】

【課題を解決するための手段】上記目的を達成するために、本発明の構造化文書圧縮装置（請求項１）は、同一のデータ構造を有する複数の構造化文書を圧縮する装置であって、該構造化文書におけるタグを出現順序に従って抽出してリスト化した該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、該タグリスト生成部により生成された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部とをそなえたことを特徴としている。

【００１６】また、本発明の構造化文書復元装置（請求項３）は、上述した本発明の構造化文書圧縮装置（請求項１）により生成された複数の圧縮文書を復元する装置であって、タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製する複製部と、その複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえたことを特徴としている。

【００１７】そして、本発明の構造化文書処理システム（請求項５）は、同一のデータ構造を有する複数の構造化文書に対する処理を行なうべく、上述した本発明の構造化文書圧縮装置（請求項１）および構造化文書復元装置（請求項３）を含んで構成されたことを特徴としている。

【００１８】上述した、本発明の構造化文書圧縮装置（請求項１）、あるいは、本発明の構造化文書処理システム（請求項５）における構造化文書圧縮装置では、複数の構造化文書について共通のデータ構造が、タグリスト取得部により、一つのタグリストとして取得されるとともに、各構造化文書は、構造化文書圧縮部のタグ圧縮処理（タグを所定の区切りコードに置き換える処理）により圧縮文書に変換された後、一つのタグリストと複数の圧縮文書とが、複数の構造化文書についての圧縮結果として出力部から出力される。

【００１９】従って、複数の構造化文書の圧縮結果は、データ構造（一つのタグリスト）とデータ内容（複数の圧縮文書）とに分離された状態で出力される。また、上述のごとく生成された圧縮文書では、タグが所定の区切りコードに置換されているだけで、データ内容（要素内容）はそのまま記述されている。このため、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を圧縮することができる。

【００２０】さらに、上述した、本発明の構造化文書復元装置（請求項３）、あるいは、本発明の構造化文書処理システム（請求項５）における構造化文書復元装置では、タグリストに対応するデータ構造が、複製部により、圧縮文書毎にメモリ上に複製データ構造として展開・複製される。そして、各圧縮文書中の要素内容が、書出部により、複製データ構造におけるタグの位置と各圧縮文書中の所定の区切りコードの位置との対応をとりながら、メモリ上における複製データ構造の所定領域に書き出される。これにより、各圧縮文書（各構造化文書）が、メモリ上に展開された状態で復元されることになる。

【００２１】このとき、複数の構造化文書の圧縮結果は、データ構造（一つのタグリスト）とデータ内容（複数の圧縮文書）とに分離されているので、一つのタグリストに対する解析処理を一度だけ行なって、複数の圧縮文書に共通のデータ構造を取得してしまえば、後は、取得されたデータ構造を複製して流用することにより、圧縮文書毎に一々タグ解析を行なう必要をなくすことができる。

【００２２】一方、本発明の構造化文書圧縮装置（請求項２）は、例えば上述した構造化文書圧縮装置（請求項１）における構造化文書圧縮部として用いられるものであって、構造化文書を圧縮すべく、該構造化文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえたことを特徴としている。

【００２３】また、本発明の構造化文書復元装置（請求項４）は、上述した本発明の構造化文書圧縮装置（請求項２）により生成された圧縮文書を復元する装置であって、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書中の所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえたことを特徴としている。

【００２４】上述した本発明の構造化文書圧縮装置（請求項２）では、構造化文書中のタグがタグ検出部により検出されると、そのタグは、タグ圧縮部により所定の区切りコードに置き換えられて圧縮される。このような単純な置換処理によって圧縮文書が生成される。そして、生成された圧縮文書では、上述した通り、タグが所定の

区切りコードに置換されているだけで、データ内容（要素内容）はそのまま記述されている。従って、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を圧縮することができる。

【００２５】また、上述した本発明の構造化文書復元装置（請求項４）では、復元対象の圧縮文書に対応したタグリストが、タグリスト保持部に予め保持されており、圧縮文書中の区切りコードが区切りコード検出部によって検出されると、その区切りコードは、タグ復元部により、その区切りコードに対応したタグに置き換えられる。

【００２６】このとき、タグ復元部においては、タグリストにおけるタグの位置と区切りコード検出部により検出された所定の区切りコードの位置との対応をとることにより、検出された所定の区切りコードがタグリスト中のどのタグに対応するかが認識されるので、所定の区切りコードを適切なタグに復元することができる。これにより、圧縮文書は、元の構造化文書に復元される。

【００２７】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を説明する。

〔１〕第１実施形態の説明

図１は本発明の第１実施形態における構造化文書圧縮装置の構成を示すブロック図、図２は本発明の第１実施形態における構造化文書復元装置（メモリ展開部）の構成を示すブロック図である。

【００２８】図１に示す構造化文書圧縮装置１１０および図２に示すメモリ展開部（構造化文書復元装置）２１０は、同一のデータ構造（文書構造，木構造）を有する複数の構造化文書に対する処理を行なうための構造化文書処理システムに含まれ、この構造化文書処理システムの一部を構成するものである。

【００２９】なお、以下に説明する各実施形態においては、構造化文書がＸＭＬ(eXtensible Markup Language)であり、特に、同一のデータ構造を有する多数のＸＭＬ文書を取り扱う場合について説明する。また、構造化文書処理システムによって実行される処理は、例えばＸＭＬ文書の蓄積，加工，転送等である。なお、処理対象となるＸＭＬ文書は、例えば図３（Ａ）を参照しながら後述するような発注伝票文書である。

【００３０】そして、構造化文書処理システムにおいては、ＸＭＬ文書のための記憶容量を削減したり前記処理を高速化すべくＸＭＬ文書を圧縮するための構造化文書圧縮装置１１０がそなえられ、さらに、構造化文書圧縮装置１１０による圧縮データをメモリ２１４上に復元・展開するためのメモリ展開部（構造化文書復元装置）２１０がそなえられている。

【００３１】第１実施形態の構造化文書圧縮装置１１０は、図１に示すように、タグリスト取得部１１１，構造化文書圧縮部１１２および出力部１１３をそなえて構成される一方、第１実施形態のメモリ展開部２１０は、図２に示すように、解析部２１１，複製部２１２および書出部２１３をそなえて構成されている。

【００３２】ここで、構造化文書圧縮装置１１０およびメモリ展開部２１０は、同一のコンピュータ上、もしくは、それぞれ異なるコンピュータ上にそなえられている。そして、構造化文書圧縮装置１１０を成すタグリスト取得部１１１，構造化文書圧縮部１１２および出力部１１３としての機能は、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。同様に、メモリ展開部２１０を成す解析部２１１，複製部２１２および書出部２１３としての機能は、コンピュータ上で所定のプログラム（構造化文書復元プログラム）を実行することにより実現されるようになっている。

【００３３】さて、図１において、ハードディスク（データベース）３００は、圧縮対象である、同一データ構造を有する複数のＸＭＬ文書（図１ではＸＭＬ文書１～３）を予め格納・保持するものである。また、ハードディスク（データベース）４００は、構造化文書圧縮装置１１０による圧縮結果を格納・保持するものである。なお、圧縮対象のＸＭＬ文書と構造化文書圧縮装置１１０による圧縮結果とを、同じハードディスク３００もしくは４００に格納するように構成してもよい。

【００３４】タグリスト取得部１１１は、ハードディスク３００に格納された複数のＸＭＬ文書について共通の一つのタグリストを取得するものである。タグリストは、図３（Ｂ）を参照しながら後述するごとく、ＸＭＬ文書におけるタグを出現順序に従って抽出してリスト化したもので、どのようなタグがどのような順序で出現するかを示すものである。同一データ構造を有する複数のＸＭＬ文書についてのタグリストは全て同一のものとなる。

【００３５】このタグリスト取得部１１１は、予め生成されデータベース（図示略）等に格納されているタグリストを、そのデータベースから取得してもよいし、ハードディスク３００に格納されている複数のＸＭＬ文書のうちの少なくとも一つから、タグリストを抽出・生成して取得してもよい。構造化文書圧縮部１１２は、各ＸＭＬ文書中のタグを所定の区切りコードに置き換えたＸＭＬ圧縮文書を生成するものである。なお、第１実施形態では、図３（Ｃ）を参照しながら後述するごとく、所定の区切りコードとして「，」を用いる。

【００３６】出力部１１３は、タグリスト取得部１１１により取得された一つのタグリストと、複数のＸＭＬ文書のそれぞれについて構造化文書圧縮部１１２により生成された複数のＸＭＬ圧縮文書（図１ではＸＭＬ圧縮文書１～３）とを対応させ複数のＸＭＬ文書の圧縮結果としてハードディスク４００に出力・格納するものであ

る。

【００３７】一方、図２において、読出部５００は、ハードディスク４００に格納された、共通のタグリストと複数のＸＭＬ圧縮文書とをメモリ展開部２１０へ読み出すものであり、メモリ展開部２１０は、タグリストに基づいて複数のＸＭＬ圧縮文書をメモリ２１４上に復元・展開すべく、解析部２１１，複製部２１２および書出部２１３を有している。

【００３８】解析部２１１は、ハードディスク４００から読出部５００によって読み出されたタグリストを解析し、復元・展開対象である複数のＸＭＬ圧縮文書について共通のデータ構造を解析結果として得るものである。複製部２１２は、解析部２１１によって得られた、タグリストに対応するデータ構造を、メモリ２１４上に複製データ構造として展開・複製するものである。書出部２１３は、複製データ構造におけるタグの位置と各ＸＭＬ圧縮文書中の区切りコード「，」の位置とを対応させながら、各ＸＭＬ圧縮文書中の要素内容をメモリ２１４上における複製データ構造の所定領域に書き出すものである。

【００３９】次に、上述のごとく構成された、第１実施形態の構造化文書圧縮装置１１０およびメモリ展開部２１０の動作について説明する。図１に示す構造化文書圧縮装置１１０においては、複数のＸＭＬ文書について共通のデータ構造が、タグリスト取得部１１１により、一つのタグリストとして取得されるとともに、各ＸＭＬ文書は、構造化文書圧縮部１１２のタグ圧縮処理により、タグを区切りコード「，」に置き換えたＸＭＬ圧縮文書に変換される。

【００４０】この後、タグリスト取得１１１により取得されたタグリストと、構造化文書圧縮部１１２により得られた複数のＸＭＬ圧縮文書とが、複数のＸＭＬ文書についての圧縮結果として出力部１１３から出力され、ハードディスク４００に格納される。つまり、第１実施形態では、複数のＸＭＬ文書の圧縮結果が、データ構造（タグ情報）とデータ内容（タグ情報以外の情報）とに分離された状態で出力されることになる。なお、データ構造（タグ情報）は、前記一つのタグリストであり、データ内容は、区切りコードと要素内容とからなる、複数のＸＭＬ圧縮文書である。

【００４１】このとき、タグリストと各ＸＭＬ圧縮文書とは、例えば図１５～図１７を参照しながら後述する手法等によって対応付けられており、複数のＸＭＬ文書について共通のデータ構造を示す一つのタグリストは、複数のＸＭＬ圧縮文書によって共有される。

【００４２】ここで、図３（Ａ）～図３（Ｃ）を参照しながら、第１実施形態における具体的なＸＭＬ文書の圧縮状態について説明する。なお、図３（Ａ）～図３（Ｃ）はいずれも第１実施形態におけるデータ例を示すもので、図３（Ａ）はＸＭＬ文書の一例を示す図、図３（Ｂ）は図３（Ａ）に示すＸＭＬ文書から得られたタグリストを示す図、図３（Ｃ）は図３（Ａ）に示すＸＭＬ文書の圧縮状態を示す図である。

【００４３】図３（Ａ）には、圧縮前つまり圧縮対象のＸＭＬ文書の一例として、発注伝票をＸＭＬにより記述した例が示されている。この図３（Ａ）に示すＸＭＬ文書では、開始タグ＜発注伝票＞，＜発注者＞，＜名前＞，＜電話番号＞，＜商品＞，＜メーカ＞，＜製品番号＞，＜製品名＞，＜価格＞と、終了タグ＜/発注伝票＞，＜/発注者＞，＜/名前＞，＜/電話番号＞，＜/商品＞，＜/メーカ＞，＜/製品番号＞，＜/製品名＞，＜/価格＞とにより、ＸＭＬ文書のデータ構造（つまり発注伝票のフォーマット）が定義されている。

【００４４】この図３（Ａ）に示すＸＭＬ文書においては、開始タグ＜名前＞と終了タグ＜/名前＞との間には、発注者の名前「ＳＴＵＶ」が要素内容として記述され、開始タグ＜電話番号＞と終了タグ＜/電話番号＞との間には、発注者の電話番号「1111」が要素内容として記述され、開始タグ＜メーカ＞と終了タグ＜/メーカ＞との間には、商品のメーカ「Ａ社」が要素内容として記述され、開始タグ＜製品番号＞と終了タグ＜/製品番号＞との間には、商品の製品番号「1234」が要素内容として記述され、開始タグ＜製品名＞と終了タグ＜/製品名＞との間には、商品の製品名「ＡＢＣＤ」が要素内容として記述され、開始タグ＜価格＞と終了タグ＜/価格＞との間には、商品の価格「980」が要素内容として記述されている。

【００４５】また、図３（Ｂ）は、図３（Ａ）に示したＸＭＬ文書のタグリストを示している。このタグリストは、前述した通り、予め何らかの手段により作成されているか、もしくは、タグリスト取得部１１１により、図３（Ａ）に示すＸＭＬ文書から直接的に抽出して作成されるもので、図３（Ｂ）に示す例では、図３（Ａ）のＸＭＬ文書から、ただ単に要素内容「ＳＴＵＶ」，「1111」，「Ａ社」，「ＡＢＣＤ」，「980」を取り除いた構成となっている。

【００４６】そして、図３（Ｃ）には、図３（Ａ）に示すＸＭＬ文書に対し、構造化文書圧縮部１１２によるタグ圧縮処理を施した結果、即ち、図３（Ａ）に示すＸＭＬ文書中のタグを区切りコード「，」に置き換えたＸＭＬ圧縮文書が示されている。これらの図３（Ａ）～図３（Ｃ）を比較対照しても明らかなように、タグリスト中の各タグとＸＭＬ圧縮文書中の各区切りコード「，」とは一対一で対応するとともに、ＸＭＬ圧縮文書において区切りコード「，」はタグの位置に対応して配置される。また、ＸＭＬ文書中の要素内容は、ＸＭＬ圧縮文書中においてそのまま記述されている。従って、第１実施形態のＸＭＬ圧縮文書は、ＸＭＬ文書と同様、自由なデータ構造を表現することができるほか、テキストで記述されるため、視認性を維持することもできる。

【００４７】一方、図２に示すメモリ展開部２１０においては、まず、復元・展開対象のＸＭＬ圧縮文書に対応付けられたタグリストが、ハードディスク４００から読出部５００により読み出されて解析部２１１に入力される。この解析部２１１においては、入力されたタグリストが解析され、その解析結果として、復元・展開対象の複数のＸＭＬ圧縮文書について共通のデータ構造が得られる。そして、解析部２１１で得られたデータ構造は、複製部２１２により、ＸＭＬ圧縮文書毎にメモリ２１４上に複製データ構造として展開・複製される。

【００４８】この後、各ＸＭＬ圧縮文書中の要素内容が、書出部２１３により、複製データ構造におけるタグの位置と、各ＸＭＬ圧縮文書中の区切りコード「，」の位置との対応をとりながら、メモリ２１４上における複製データ構造の所定領域に書き出される。これにより、各ＸＭＬ圧縮文書（各構造化文書）が、メモリ２１４上に展開された状態で復元されることになる。

【００４９】このように、本発明の第１実施形態によれば、構造化文書圧縮部１１２により生成された各ＸＭＬ圧縮文書においては、タグが区切りコード「，」に置換されているだけで、データ内容（要素内容）はそのまま記述されているので、ＸＭＬ文書（構造化文書）の利点であるデータ構造の視認性や柔軟性／拡張性を生かしたまま、ＸＭＬ文書を圧縮してＸＭＬ文書のデータ量を削減することができる。

【００５０】従って、ＸＭＬ文書（ＸＭＬ圧縮文書）を格納するための記憶領域の容量を削減することができ、ＸＭＬ圧縮文書を格納する記憶媒体（本実施形態ではハードディスク４００）の記憶領域を有効に利用できるようになるほか、ＸＭＬ文書データの伝送速度を高速化することができる。

【００５１】また、複数のＸＭＬ文書の圧縮結果は、データ構造（一つのタグリスト）とデータ内容（複数のＸＭＬ圧縮文書）とに分離されているので、第１実施形態のメモリ展開部２１０では、一つのタグリストに対する解析処理を解析部２１１において一度だけ行ない、複数のＸＭＬ圧縮文書に共通のデータ構造を取得してしまえば、後は、取得されたデータ構造を複製部２１２により複製して流用することで、ＸＭＬ圧縮文書毎に一々タグ解析を行なう必要がなくなる。

【００５２】これにより、同一のデータ構造を有する多数のＸＭＬ文書を取り扱う際に、メモリ展開部２１０の解析部２１１において無駄なタグ解析を行なう必要が一切なくなるので、タグ解析の負荷が大幅に低減され、ＸＭＬ文書をメモリ２１４に展開する際の処理速度を飛躍的に高速化することができる。

【００５３】〔２〕第２実施形態の説明

図４は本発明の第２実施形態における構造化文書圧縮装置の構成を示すブロック図、図５は本発明の第２実施形態における構造化文書復元装置の構成を示すブロック図である。図４に示す構造化文書圧縮装置１２０および図５に示す構造化文書復元装置２２０は、ＸＭＬ文書に対する処理を行なうための構造化文書処理システムに含まれて、この構造化文書処理システムの一部を構成するものである。

【００５４】この第２実施形態の構造化文書圧縮装置１２０は、ＸＭＬ文書を圧縮するためのもので、図４に示すように、入力部１２１，タグ検出部１２２，タグ圧縮部１２３および出力部１２４をそなえて構成されている。なお、構造化文書圧縮装置１２０は、第１実施形態の構造化文書圧縮部１１２として用いることも可能である。

【００５５】また、第２実施形態の構造化文書復元装置２２０は、構造化文書圧縮装置１２０により生成されたＸＭＬ圧縮文書（圧縮データ）をＸＭＬ文書に復元するためのもので、図５に示すように、入力部２２１，タグリスト保持部２２２，区切りコード検出部２２３，タグ復元部２２４および出力部２２５をそなえて構成されている。

【００５６】ここで、構造化文書圧縮装置１２０および構造化文書復元装置２２０は、同一のコンピュータ上、もしくは、それぞれ異なるコンピュータ上にそなえられている。そして、構造化文書圧縮装置１２０を成す入力部１２１，タグ検出部１２２，タグ圧縮部１２３および出力部１２４としての機能は、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。同様に、構造化文書復元装置２２０を成す入力部２２１，区切りコード検出部２２３，タグ復元部２２４および出力部２２５としての機能は、コンピュータ上で所定のプログラム（構造化文書復元プログラム）を実行することにより実現されるようになっている。

【００５７】さて、図４に示す構造化文書圧縮装置１２０において、入力部１２１は、圧縮対象のＸＭＬ文書を、ハードディスク等（例えば図１の符号３００参照）から取り込むものであり、タグ検出部１２２は、入力部１２１により取り込まれたＸＭＬ文書中のタグを検出するものである。

【００５８】タグ圧縮部１２３は、タグ検出部１２２により検出されたタグを、所定の区切りコードに置き換えて圧縮するものである。なお、第２実施形態では、第１実施形態と同様、図６（Ｃ）を参照しながら後述するごとく、所定の区切りコードとして「，」を用いる。また、２種類の区切りコード「，」および「/」を準備しておき、タグ圧縮部１２３が、これら２種類の区切りコードを開始タグと終了タグとで使い分け、図６（Ｄ）を参照しながら後述するごとく、開始タグを「，」に置き換えるとともに終了タグを「/」に置き換えるように構成してもよい。出力部１２４は、タグ圧縮部１２３を用いて生成されたＸＭＬ圧縮文書を、圧縮結果として、ハ

ードディスク等（例えば図１，図２，図１５～図１７の符号４００，４１０，４２０，４４０参照）に出力・格納するものである。

【００５９】一方、図５に示す構造化文書復元装置２２０において、入力部２２１は、復元対象のＸＭＬ圧縮文書を、記憶媒体等（例えば図１，図２，図１５～図１７に示すハードディスク４００，４１０，４２０，４４０）から取り込むものである。タグリスト保持部２２２は、ＸＭＬ文書におけるタグを出現順序に従ってリスト化したタグリスト〔例えば図６（Ｂ）参照〕を予め保持するものである。このタグリスト保持部２２２には、予め生成されたタグリストをデータベース（図示略）等から取得して格納する。

【００６０】なお、第２実施形態においても、第１実施形態と同様、処理対象となる複数のＸＭＬ文書が同一のデータ構造を有していることを前提としており、タグリストは、第１実施形態において前述した通り、これら複数のＸＭＬ文書により共有され、各ＸＭＬ文書において、どのようなタグがどのような順序で出現するかを示すものである。

【００６１】区切りコード検出部２２３は、入力部２２１により取り込まれたＸＭＬ圧縮文書中の区切りコードを検出するものである。タグ復元部２２４は、タグリスト保持部２２２に保持されたタグリストにおけるタグの位置と、区切りコード検出部２２３により検出された区切りコードの位置とを対応させながら、その区切りコードを、タグリストにおける対応するタグに置き換えて復元するものである。出力部２２５は、タグ復元部２２４を用いて復元されたＸＭＬ文書を、復元結果として、記憶媒体等（例えば図１に示すハードディスク３００）に出力・格納するものである。

【００６２】次に、上述のごとく構成された、第２実施形態の構造化文書圧縮装置１２０および構造化文書復元装置２２０の動作について説明する。図４に示す構造化文書圧縮装置１２０においては、まず、圧縮対象のＸＭＬ文書を入力部１２１により取り込み、そのＸＭＬ文書中のタグをタグ検出部１２２により探索する。タグ以外の部分（つまり要素内容の部分）はそのまま出力部１２４へ送られるが、タグ検出部１２２によりタグが検出されると、そのタグは、タグ圧縮部１２３により所定の区切りコードに置き換えられて圧縮されてから、出力部１２４へ送られる。このような単純な置換処理によって生成されたＸＭＬ圧縮文書が、圧縮結果として出力部１２４から出力される。

【００６３】ここで、図６（Ａ）～図６（Ｄ）を参照しながら第２実施形態における具体的なＸＭＬ文書の圧縮状態について説明する。なお、図６（Ａ）～図６（Ｄ）はいずれも第２実施形態におけるデータ例を示すもので、図６（Ａ）はＸＭＬ文書の一例を示す図、図６（Ｂ）は図６（Ａ）に示すＸＭＬ文書に対応するタグリストを示す図、図６（Ｃ）は図６（Ａ）に示すＸＭＬ文書の圧縮状態の一例を示す図、図６（Ｄ）は図６（Ａ）に示すＸＭＬ文書の圧縮状態の他例を示す図である。

【００６４】図６（Ａ）には、圧縮前つまり圧縮対象のＸＭＬ文書の一例として、発注伝票をＸＭＬにより記述した例が示されている。特に、図６（Ａ）では、図３（Ａ）を参照しながら前述した発注伝票の一部分（商品のメーカ，製品番号および価格にかかる記述部分）が抽出されて示されている。

【００６５】また、図６（Ｂ）は、図６（Ａ）に示したＸＭＬ文書のタグリストを示しており、このようなタグリストが、予め何らかの手段により抽出・生成されて、構造化文書復元装置２２０のタグリスト保持部２２２に格納されている。なお、図６（Ｂ）に示すタグリストでは、タグの前後に付される括弧表示（“＜”および“＞”）が省略されている。

【００６６】そして、図６（Ｃ）には、図６（Ａ）に示すＸＭＬ文書に対し、タグ圧縮部１２３によるタグ圧縮処理を施した結果、即ち、図６（Ａ）に示すＸＭＬ文書中のタグを区切りコード「，」に置き換えたＸＭＬ圧縮文書が示されている。また、図６（Ｄ）には、同一のＸＭＬ文書についての他の圧縮結果が示されている。つまり、図６（Ｄ）に示すＸＭＬ圧縮文書は、タグ圧縮部１２３によるタグ圧縮処理に際して、ＸＭＬ文書中の開始タグを「，」に置き換え、ＸＭＬ文書中の終了タグを「/」に置き換えたものである。

【００６７】これらの図６（Ａ）～図６（Ｄ）を比較対照しても明らかなように、タグリスト中の各タグとＸＭＬ圧縮文書中の各区切りコード「，」または「/」とは一対一で対応するとともに、ＸＭＬ圧縮文書において区切りコード「，」または「/」はタグの位置に対応して配置される。また、ＸＭＬ文書中の要素内容は、ＸＭＬ圧縮文書中においてそのまま記述されている。

【００６８】従って、第２実施形態のＸＭＬ圧縮文書によっても、ＸＭＬ文書と同様の自由なデータ構造表現が可能であり、要素内容の視認性が維持される。特に、図６（Ｄ）に示すＸＭＬ圧縮文書では、２種類の区切りコード「，」と「/」とがそれぞれ開始タグと終了タグとに対応して用いられるので、開始タグおよび終了タグの位置を視認することも可能になる。

【００６９】一方、図５に示す構造化文書復元装置２２０においては、まず、例えば図６（Ｃ）もしくは図６（Ｄ）に示すような復元対象のＸＭＬ圧縮文書を入力部２２１により取り込み、そのＸＭＬ圧縮文書中の区切りコード（「，」もしくは「，」と「/」）を区切りコード検出部２２３により探索する。

【００７０】区切りコード以外の部分（つまり要素内容の部分）は、そのまま出力部２２５へ送られるが、区切りコード検出部２２３により区切りコードが検出されると、その区切りコードは、タグ復元部２２４により、そ

の区切りコードに対応したタグに置き換えられてから、出力部２２５へ送られる。このような単純な置換処理によって例えば図６（Ａ）に示すようなＸＭＬ文書が復元され、出力部２２５から出力される。

【００７１】タグ復元部２２４による置換処理に際しては、タグリスト保持部２２２に保持されたタグリストにおけるタグの位置と、区切りコード検出部２２３により検出された区切りコードの位置との対応をとることにより、検出された区切りコードがタグリスト中のどのタグに対応するかが認識されるので、区切りコードを適切なタグ（対応するタグ）に復元することができる。このようにして、ＸＭＬ圧縮文書は、元のＸＭＬ文書に復元される。

【００７２】このように、本発明の第２実施形態の構造化文書圧縮装置１２０によれば、ＸＭＬ文書中で検出されたタグを所定の区切りコードに置換するという極めて単純な置換処理によって、ＸＭＬ文書（構造化文書）の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、ＸＭＬ文書を圧縮してＸＭＬ文書のデータ量を削減することができる。

【００７３】従って、第２実施形態においても、第１実施形態と同様、ＸＭＬ文書（ＸＭＬ圧縮文書）を格納するための記憶領域の容量を削減することができ、ＸＭＬ圧縮文書を格納する記憶媒体（例えば図１，図２，図１５～図１７に示すハードディスク４００，４１０，４２０，４４０）の記憶領域を有効に利用できるようになるほか、ＸＭＬ文書データの伝送速度を高速化することができる。

【００７４】また、第２実施形態の構造化文書復元装置２２０によれば、ＸＭＬ圧縮文書中で検出された区切りコードを、そのＸＭＬ圧縮文書についてのタグリスト中のタグと対応させながら、所定のタグに置き換えるという簡易な置換処理によって、ＸＭＬ圧縮文書を極めて容易に元のＸＭＬ文書に復元することができるという利点もある。

【００７５】〔３〕第３実施形態の説明

図７は本発明の第３実施形態における構造化文書圧縮装置の構成を示すブロック図、図８は本発明の第３実施形態における構造化文書復元装置の構成を示すブロック図である。なお、図中、既述の符号と同一の符号は同一の部分もしくはほぼ同一の部分を示しているので、その詳細な説明は省略する。

【００７６】図７に示す構造化文書圧縮装置１３０および図８に示す構造化文書復元装置２３０も、第２実施形態と同様、ＸＭＬ文書に対する処理を行なうための構造化文書処理システムに含まれて、この構造化文書処理システムの一部を構成するもので、それぞれ、図４に示す構造化文書圧縮装置１２０および図５に示す構造化文書復元装置２２０とほぼ同様に構成されている。

【００７７】ただし、第３実施形態の構造化文書圧縮装置１３０は、ＸＭＬ文書のタグ内に属性が記述されている場合には、その属性を圧縮後も残すことにより属性の視認性をも維持しながら、ＸＭＬ文書の圧縮を行なえるように構成したもので、図７に示すように、第２実施形態と同様の入力部１２１，タグ検出部１２２，タグ圧縮部１２３および出力部１２４のほか、さらに、属性付きタグ検出部１３１および属性付きタグ圧縮部１３２をそなえて構成されている。なお、この構造化文書圧縮装置１３０も、第１実施形態の構造化文書圧縮部１１２として用いることが可能である。

【００７８】また、第３実施形態の構造化文書復元装置２３０は、構造化文書圧縮装置１３０により生成されたＸＭＬ圧縮文書（圧縮データ）をＸＭＬ文書に復元するためのもので、図８に示すように、第２実施形態と同様の入力部２２１，タグリスト保持部２２２，区切りコード検出部２２３，タグ復元部２２４および出力部２２５のほか、さらに、属性リスト保持部２３１，属性付きタグ検出部２３２および属性付きタグ復元部２３３をそなえて構成されている。

【００７９】ここで、第３実施形態の構造化文書圧縮装置１３０および構造化文書復元装置２３０も、同一のコンピュータ上、もしくは、それぞれ異なるコンピュータ上にそなえられている。そして、構造化文書圧縮装置１３０を成す入力部１２１，タグ検出部１２２，タグ圧縮部１２３，出力部１２４，属性付きタグ検出部１３１および属性付きタグ圧縮部１３２としての機能は、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。同様に、構造化文書復元装置２３０を成す入力部２２１，区切りコード検出部２２３，タグ復元部２２４，出力部２２５，属性付きタグ検出部２３２および属性付きタグ復元部２３３としての機能は、コンピュータ上で所定のプログラム（構造化文書復元プログラム）を実行することにより実現されるようになっている。

【００８０】さて、図７に示す構造化文書圧縮装置１３０において、属性付きタグ検出部１３１は、タグ検出部１２２に含まれており、タグ検出部１２２により検出されたタグが属性値をもつ属性付きタグであるか否かを検出するものである。なお、属性付きタグとは、そのタグ内に、要素内容に付加したい情報（属性）を記述されたものである。その属性は、具体的には図９（Ａ）を参照しながら後述するごとく、開始タグ内において、要素名の後にスペースを空け「属性名＝‘属性値’」として記述される。つまり、属性付きタグは、一般的には「＜要素名　属性名＝‘属性値’＞」と記述される。

【００８１】属性付きタグ圧縮部１３２は、属性付きタグ検出部１３１により検出された属性付きタグを、そのタグ内に記述された属性値と所定の区切りコードとにより置き換えて圧縮するものである。この属性付きタグ圧縮部１３２によって置き換えられる区切りコードとして

は、例えば図９（Ｃ）や図９（Ｄ）を参照しながら後述するごとく「，」あるいは「=」を用いる。

【００８２】また、本実施形態では、属性付きタグ圧縮部１３２により属性付きタグを属性値と区切りコードとに置き換える際、その区切りコードは属性値の前後に配置されるようになっている〔図９（Ｃ）や図９（Ｄ）参照〕。例えば「＜要素名　属性名＝‘属性値’＞」は「，属性値，」もしくは「，属性値=」という圧縮記述に置き換えられる。また、複数の属性をもつタグ、例えば「＜要素名　属性名１＝‘属性値１’属性名２＝‘属性値２’＞」と記述されたタグは「，属性値１，属性値２，」もしくは「，属性値１=属性値２=」という圧縮記述に置き換えられる。

【００８３】なお、第３実施形態の出力部１２４は、タグ圧縮部１２３および属性付きタグ圧縮部１３２を用いて生成されたＸＭＬ圧縮文書を、圧縮結果として、ハードディスク等（例えば図１，図２，図１５～図１７の符号４００，４１０，４２０，４４０参照）に出力・格納するようになっている。

【００８４】一方、図８に示す構造化文書復元装置２３０において、属性リスト保持部２３１は、ＸＭＬ圧縮文書における属性名を出現順序に従ってリスト化した属性リストを予め保持するものである。この属性リスト保持部２３１には、予め生成された属性リストをデータベース（図示略）等から取得して格納する。

【００８５】ここで、第３実施形態の属性リストは、実際には、図９（Ｂ）を参照しながら後述するごとく、タグリストに含まれる形で与えられるものである。このため、図８では、属性リスト保持部２３１がタグリスト保持部２２２に含まれている。以下では、属性リストを含むタグリストのことをタグ・属性リストと表記する場合がある。このようなタグ・属性リストにおいては、そのリストに記入された文字列が属性名である場合、そのことが明確に分かるように、例えば図９（Ｂ）に示すごとく、その文字列の前（左側）に、例えばコード「=」を付与している。

【００８６】また、第３実施形態においても、第１実施形態と同様、処理対象となる複数のＸＭＬ文書が同一のデータ構造を有していることを前提としており、タグリストおよび属性リスト（タグ・属性リスト）は、これら複数のＸＭＬ文書により共有され、各ＸＭＬ文書において、どのようなタグがどのような順序で出現するか、あるいは、どのような属性がどのような順序で出現するかを示すものである。

【００８７】属性付きタグ検出部２３２は、タグ復元部２２４に含まれており、タグ復元部２２４で復元対象となったタグが属性付きタグに復元されるべきものであるか否かを検出するものである。このとき、属性付きタグ検出部２３２は、区切りコードの配置状態や区切りコードの種類を認識することにより、もしくは、ＸＭＬ圧縮文書内の属性値とタグ・属性リスト内の属性名との対応関係を参照・認識することにより、復元対象のタグが、属性付きタグに復元されるべきもの、即ち、属性をもつものであるか否かを検出している。

【００８８】属性付きタグ復元部２３３は、属性付きタグ検出部２３２により復元対象として検出されたタグを、そのタグに対応した属性を有する属性付きタグに復元するものである。第３実施形態においては、復元対象となるＸＭＬ圧縮文書のうち属性付きタグに対応する部分は、まず、タグ復元部２２４において要素名のみを含む通常のタグ「＜要素名＞」として復元される。第３実施形態の属性付きタグ復元部２３３は、属性付きタグについての属性値と属性リストにおける属性名とを対応させて、属性付きタグ内の属性を復元するものである。

【００８９】より具体的に説明すると、属性付きタグ復元部２３３は、復元すべき属性に対応する属性名を属性リスト（タグ・属性リスト）から読み出し、復元すべき属性に相当する区切りコードとこの区切りコードに組み合わされたデータ（属性値）とを通常の属性記述に置き換えることで、タグ復元部２２４で復元されたタグ内に属性を復元させ、属性付きタグの復元を行なうようになっている。例えば「属性値，」または「属性値=」という属性の圧縮記述は「属性名＝‘属性値’」に置き換えられ、「属性値１，属性値２，」または「属性値１=属性値２=」という属性の圧縮記述は「属性名１＝‘属性値１’属性名２＝‘属性値２’」に置き換えられる。

【００９０】なお、第３実施形態の出力部２２５は、タグ復元部２２４および属性付きタグ復元部２３３を用いて復元されたＸＭＬ文書を、復元結果として、記憶媒体等（例えば図１に示すハードディスク３００）に出力・格納するようになっている。次に、上述のごとく構成された、第３実施形態の構造化文書圧縮装置１３０および構造化文書復元装置２３０の動作について説明する。

【００９１】図７に示す構造化文書圧縮装置１３０においては、まず、圧縮対象のＸＭＬ文書を入力部１２１により取り込み、そのＸＭＬ文書中のタグをタグ検出部１２２により探索する。タグ以外の部分（つまり要素内容の部分）は、そのまま出力部１２４へ送られるが、タグ検出部１２２によりタグが検出されると、属性付きタグ検出部１３１により、そのタグが属性付きタグであるか否かが検出される。

【００９２】属性付きタグでない場合、第２実施形態で説明した通り、そのタグは、タグ圧縮部１２３により所定の区切りコードに置き換えられて圧縮されてから、出力部１２４へ送られる。一方、属性付きタグである場合、そのタグは、属性付きタグ圧縮部１３２により、そのタグ内に記述された属性値と所定の区切りコードとに置き換えられて圧縮されてから、出力部１２４へ送られる。

【００９３】第３実施形態においては、ＸＭＬ文書が属

性付きタグを有している場合、上述のような単純な置換処理により、要素内容とともに属性値を残したままのＸＭＬ圧縮文書が生成され圧縮結果として出力部１２４から出力される。ここで、図９（Ａ）～図９（Ｄ）を参照しながら、第３実施形態における具体的なＸＭＬ文書の圧縮状態について説明する。なお、図９（Ａ）～図９（Ｄ）はいずれも第３実施形態におけるデータ例を示すもので、図９（Ａ）はＸＭＬ文書の一例を示す図、図９（Ｂ）は図９（Ａ）に示すＸＭＬ文書に対応するタグ・属性リストを示す図、図９（Ｃ）は図９（Ａ）に示すＸＭＬ文書の圧縮状態の一例を示す図、図９（Ｄ）は図９（Ａ）に示すＸＭＬ文書の圧縮状態の他例を示す図である。

【００９４】図９（Ａ）には、圧縮前つまり圧縮対象のＸＭＬ文書の一例として、発注伝票をＸＭＬにより記述した例が示されている。特に、図９（Ａ）では、図６（Ａ）に示した例とほぼ同様の発注伝票の一部分が抽出されて示されている。この図９（Ａ）に示す例では、さらに、製品番号を要素名としてもつ開始タグが属性を有している。即ち、その開始タグ（属性付きタグ）内には、属性として「製品名＝‘ＡＢＣＤ’色＝‘青’」が記述されている。ここで、「製品名」および「色」が属性名であり、「ＡＢＣＤ」および「青」が属性値である。

【００９５】また、図９（Ｂ）は、図９（Ａ）に示したＸＭＬ文書のタグ・属性リストを示しており、このようなタグ・属性リストが、予め何らかの手段により抽出・生成されて、構造化文書復元装置２３０のタグリスト保持部２２２および属性リスト保持部２３１に格納されている。この図９（Ｂ）に示すタグ・属性リストは、図６（Ｂ）に示したタグリストに、製品番号の属性名に係る項目、つまり「=製品名」および「=色」をさらに追加したものである。

【００９６】そして、図９（Ｃ）には、図９（Ａ）に示すＸＭＬ文書に対し、タグ圧縮部１２３および属性付きタグ圧縮部１３２による圧縮処理を施した結果、即ち、図９（Ａ）に示すＸＭＬ文書中のタグを区切りコード「，」に置き換えるとともに、属性を“属性値＋区切りコード「，」”に置き換えたＸＭＬ圧縮文書が示されている。つまり、図９（Ａ）における属性付きタグ「＜製品番号　製品名＝‘ＡＢＣＤ’色＝‘青’＞」は、図９（Ｃ）に示すＸＭＬ圧縮文書では、「，ＡＢＣＤ，青，」に置き換えられている。

【００９７】また、図９（Ｄ）には、同一のＸＭＬ文書についての他の圧縮結果が示されている。つまり、図９（Ｄ）に示すＸＭＬ圧縮文書は、タグ圧縮部１２３によるタグ圧縮処理に際して、ＸＭＬ文書中の開始タグを「，」に置き換え、ＸＭＬ文書中の終了タグを「/」に置き換えるとともに、属性付きタグ圧縮部１３２による圧縮処理に際し、属性値に付加する区切りコードとして「=」を用いたものである。従って、図９（Ａ）における属性付きタグ「＜製品番号　製品名＝‘ＡＢＣＤ’色＝‘青’＞」は、図９（Ｄ）に示すＸＭＬ圧縮文書では、「，ＡＢＣＤ=青=」に置き換えられている。

【００９８】これらの図９（Ａ）～図９（Ｄ）を比較対照しても明らかなように、タグ・属性リスト中の各タグとＸＭＬ圧縮文書中の各区切りコード「，」または「/」とは一対一で対応するとともに、ＸＭＬ圧縮文書において区切りコード「，」または「/」はタグの位置に対応して配置される。また、ＸＭＬ文書中の要素内容は、ＸＭＬ圧縮文書中においてそのまま記述されている。さらに、ＸＭＬ文書中の属性値は、ＸＭＬ圧縮文書中において、右側に区切りコード「，」または「=」を付加された状態で、そのまま記述されている。

【００９９】従って、第３実施形態のＸＭＬ圧縮文書によっても、ＸＭＬ文書と同様の自由なデータ構造表現が可能であり、要素内容のみならず属性値についても視認性が維持される。特に、図９（Ｄ）に示すＸＭＬ圧縮文書では、３種類の区切りコード「，」と「/」と「=」がそれぞれ開始タグと終了タグと属性とに対応して用いられるので、開始タグ，終了タグおよび属性（属性付きタグ）の位置を視認することも可能になる。

【０１００】一方、図８に示す構造化文書復元装置２３０においては、まず、例えば図９（Ｃ）もしくは図９（Ｄ）に示すような復元対象のＸＭＬ圧縮文書を入力部２２１により取り込み、そのＸＭＬ圧縮文書中の区切りコード（「，」や「/」）を区切りコード検出部２２３により探索する。

【０１０１】区切りコードおよび属性値以外の部分（つまり要素内容の部分）は、そのまま出力部２２５へ送られるが、区切りコード検出部２２３により区切りコードが検出されると、その区切りコードは、タグ復元部２２４により、その区切りコードに対応したタグに置き換えられる。タグ復元部２２４による置換処理に際しては、第２実施形態と同様、タグリスト保持部２２２に保持されたタグリストにおけるタグの位置と、区切りコード検出部２２３により検出された区切りコードの位置との対応をとることにより、検出された区切りコードがタグリスト中のどのタグに対応するかが認識されるので、区切りコードを適切なタグ（対応するタグ）に復元することができる。

【０１０２】そして、第３実施形態では、属性付きタグ検出部２３２により、タグ復元部２２４で復元対象となったタグが属性をもつものであるか否かを検出し、属性をもたないものであると認識された場合には、タグ復元部２２４で復元されたタグ（属性をもたないタグ）は、そのまま出力部２２５へ送られる。一方、属性をもつものであると認識された場合には、タグ復元部２２４により要素名のみを含む状態で復元された通常のタグ（例えば＜製品番号＞）内に、そのタグに対応する属性を、属

性付きタグ復元部２３３によって復元させてから、出力部２２５へ送られる。

【０１０３】例えば図９（Ｃ）や図９（Ｄ）に示す圧縮記述「，ＡＢＣＤ，青，」や「，ＡＢＣＤ=青=」については、その圧縮記述の最初の区切りコード「，」が検出され、その区切りコードが「製品番号」に対応するものであることが認識される。さらに、図９（Ｂ）に示すタグ・属性リストを参照することにより、上記圧縮記述に対応するタグは、「製品名」および「色」を属性名とする２つの属性をもつことが認識される。このような属性情報の認識に応じて、属性付きタグ復元部２３３により、上記圧縮記述は、図９（Ａ）に示すような属性付きタグ「＜製品番号製品名＝‘ＡＢＣＤ’色＝‘青’＞」に変換・復元される。

【０１０４】このように、本発明の第３実施形態の構造化文書圧縮装置１３０によれば、第２実施形態の構造化文書圧縮装置１２０と同様の作用効果が得られるほか、タグが属性値をもつ属性付きタグである場合には、その属性付きタグが属性値および所定の区切りコードに置き換えられて圧縮される。これにより、ＸＭＬ圧縮文書において属性値がそのまま記述されるので、要素内容だけでなく属性値の視認性を保ちながらＸＭＬ文書の圧縮を行なうことができる。

【０１０５】また、第３実施形態の構造化文書復元装置２３０によれば、第２実施形態の構造化文書圧縮装置２２０と同様の作用効果が得られるほか、上述のような圧縮を施された属性付きタグが復元対象になると、その属性付きタグについての属性値とＸＭＬ圧縮文書についてのタグ・属性リスト中の属性名とを対応させることにより、属性付きタグを極めて容易に復元することができる。

【０１０６】〔４〕第４実施形態の説明

図１０は本発明の第４実施形態における構造化文書圧縮装置の要部構成を示すブロック図である。この図１０に示す構造化文書圧縮装置１４０は、図４に示す構造化文書圧縮装置１２０の前段に、さらに、入力部１４１，タグリスト保持部１４２，タグ並び替え部１４３および省略タグ補完部１４４をそなえて構成されたものである。なお、この構造化文書圧縮装置１４０も、第１実施形態の構造化文書圧縮部１１２として用いることが可能である。また、構造化文書圧縮装置１４０の要部を成す入力部１４１，タグ並び替え部１４３および省略タグ補完部１４４も、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。

【０１０７】さて、図１０に示す構造化文書圧縮装置１４０において、入力部１４１は、圧縮対象のＸＭＬ文書を、ハードディスク等（例えば図１の符号３００参照）から取り込むものである。また、タグリスト保持部１４２は、所定のデータ構造を定義すべく所定の順序でタグを並べたタグリストを予め保持するものである。より詳細に説明すると、第３実施形態においても、第１および第２実施形態と同様、処理対象となる複数のＸＭＬ文書が、同一のデータ構造を有していることを前提としている。そして、タグリスト保持部１４２に保持されるタグリストは、構造化文書圧縮装置２２０のタグリスト保持部２２２に保持されるタグリストと同様、これら複数のＸＭＬ文書により共有され、各ＸＭＬ文書において、どのようなタグがどのような順序で出現するかを示すものである。なお、タグリスト保持部１４２には、圧縮処理対象となるＸＭＬ文書について予め生成されたタグリストが、データベース（図示略）等から取得して格納される。

【０１０８】タグ並び替え部１４３は、入力されたＸＭＬ文書とタグリストとを比較し、ＸＭＬ文書におけるタグの記述順序をタグリストにおけるタグの配列順序（所定の順序）に合わせるように、圧縮前のＸＭＬ文書のタグを並び替えるものである。このとき、対になる開始タグと終了タグとの順序を変更する場合、タグ並び替え部１４３は、これらの開始タグと終了タグとの間に記述された要素内容も一緒に移動させる。

【０１０９】省略タグ補完部１４４は、タグリスト保持部１４２に保持されたタグリストに従って、タグ並び替え部１４３による処理を施されたＸＭＬ文書中で省略されているタグを補完するものである。つまり、省略タグ補完部１４４は、入力されたＸＭＬ文書とタグリストとを比較し、そのＸＭＬ文書中から欠落しているタグを検出すると、欠落タグに対応するタグをタグリストから読み出し、その欠落タグを補完するものである。このとき、対になる開始タグと終了タグとを補完する場合、省略タグ補完部１４４は、これらの開始タグと終了タグとの間に記述されるべき要素内容を空のままとする。

【０１１０】そして、タグ並び替え部１４３および省略タグ補完部１４４による処理を施されたＸＭＬ文書は、第２実施形態の構造化文書圧縮装置１２０に入力されるようになっている。

【０１１１】次に、上述のごとく構成された、第４実施形態の構造化文書圧縮装置１４０の動作について説明する。図１０に示す構造化文書圧縮装置１４０においては、まず、圧縮対象のＸＭＬ文書を入力部１４１により取り込み、タグ並び替え部１４３において、そのＸＭＬ文書とタグリストとが比較され、万一、ＸＭＬ文書中にタグの記述順序の逆転等の不備がある場合には、ＸＭＬ文書におけるタグの記述順序がタグリストにおけるタグの配列順序に合うように圧縮前のＸＭＬ文書のタグが並び替えられる。

【０１１２】そして、並び替え処理を施されたＸＭＬ文書は、省略タグ補完部１４４に入力され、この省略タグ補完部１４４において、そのＸＭＬ文書とタグリストとが比較され、そのＸＭＬ文書中から欠落しているタグが

検出されると、欠落タグに対応するタグがタグリストから読み出され、その欠落タグが補完される。

【０１１３】ここで、図１１（Ａ）～図１１（Ｃ）を参照しながら、第４実施形態における具体的なＸＭＬ文書の圧縮状態について説明する。なお、図１１（Ａ）～図１１（Ｃ）はいずれも第４実施形態におけるデータ例を示すもので、図１１（Ａ）はタグリストの一例を示す図、図１１（Ｂ）はタグの記述に不備のあるＸＭＬ文書の一例を示す図、図１１（Ｃ）は図１１（Ｂ）に示すＸＭＬ文書を図１１（Ａ）に示すタグリストに従って正規化した結果を示す図である。

【０１１４】図１１（Ａ）には、図６（Ｂ）に示したものと全く同じタグリストが示されており、ここでは、この図１１（Ａ）に示すタグリストに従って、タグ並び替え部１４３および省略タグ補完部１４４による処理をＸＭＬ文書に施す場合について説明する。その処理対象になるＸＭＬ文書は、例えば図１１（Ｂ）に示すものである。

【０１１５】まず、タグ並び替え部１４３において、図１１（Ａ）のタグリストと図１１（Ｂ）のＸＭＬ文書とを比較することにより、図１１（Ｂ）のＸＭＬ文書では、価格についてのタグおよび要素内容「300」と、製品番号についてのタグおよび要素内容「Ｂ７」との配置順序が逆転していることが認識され、その順序が並び替えられる。

【０１１６】そして、省略タグ補完部１４４において、上述のごとく順序を並び替えられたＸＭＬ文書と図１１（Ａ）のタグリストとを比較することにより、そのＸＭＬ文書では、メーカについてのタグが欠落していることが認識され、メーカについての開始タグと終了タグとが空要素の状態で補完される。その結果、図１１（Ｂ）に示すようにタグの記述に不備のあったＸＭＬ文書が、図１１（Ｂ）に示すタグリストに応じたデータ構造をもつＸＭＬ文書に修正され、図１１（Ｃ）に示すようなＸＭＬ文書に正規化（整頓）される。

【０１１７】つまり、構造化文書圧縮装置１２０による圧縮対象である、全てのＸＭＬ文書に対し、タグ並び替え部１４３および省略タグ補完部１４４による処理を施すことによって、全てのＸＭＬ文書が、タグリストに応じたデータ構造をもつＸＭＬ文書となるように正規化される。

【０１１８】そして、上述のごとく正規化されたＸＭＬ文書が、構造化文書圧縮装置１２０に入力され、第２実施形態で前述したように圧縮される。なお、当然、不備のないＸＭＬ文書は、タグ並び替え部１４３および省略タグ補完部１４４をそのまま通過して、構造化文書圧縮装置１２０に入力される。

【０１１９】ところで、図１２は本発明の第４実施形態における構造化文書圧縮装置の変形例の要部構成を示すブロック図である。この図１２に示す構造化文書圧縮装置１５０は、図７に示す構造化文書圧縮装置１３０の前段に、さらに、入力部１５１，タグ・属性リスト保持部１５２，タグ・属性並び替え部１５３および省略タグ・属性補完部１５４をそなえて構成されたものである。なお、この構造化文書圧縮装置１５０も、第１実施形態の構造化文書圧縮部１１２として用いることが可能である。また、構造化文書圧縮装置１５０の要部を成す入力部１５１，タグ・属性並び替え部１５３および省略タグ・属性補完部１５４も、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。

【０１２０】さて、図１２に示す構造化文書圧縮装置１５０において、入力部１５１は、属性付きタグを含む圧縮対象のＸＭＬ文書を、ハードディスク等（例えば図１の符号３００参照）から取り込むものである。また、タグ・属性リスト保持部１５２は、所定のデータ構造を定義すべく所定の順序で並べたタグと属性名とをもつタグ・属性リストを予め保持するものである。より詳細に説明すると、この第４実施形態の変形例においても、第１～第３実施形態と同様、処理対象となる複数のＸＭＬ文書が、同一のデータ構造を有していることを前提としている。そして、タグ・属性リスト保持部１５２に保持されるタグ・属性リストは、構造化文書圧縮装置２３０のタグ・属性リストと同様、これら複数のＸＭＬ文書により共有され、各ＸＭＬ文書において、どのようなタグがどのような順序で出現するか、あるいは、どのような属性がどのような順序で出現するかを示すものである。なお、タグ・属性リスト保持部１５２には、圧縮処理対象となるＸＭＬ文書について予め生成されたタグ・属性リストが、データベース（図示略）等から取得して格納される。

【０１２１】タグ・属性並び替え部１５３は、入力されたＸＭＬ文書とタグ・属性リストとを比較し、ＸＭＬ文書におけるタグおよび属性の記述順序をタグ・属性リストにおけるタグおよび属性の配列順序（所定の順序）に合わせるように、圧縮前のＸＭＬ文書のタグや属性を並び替えるものである。このとき、対になる開始タグと終了タグとの順序を変更する場合、タグ・属性並び替え部１５３は、これらの開始タグと終了タグとの間に記述された要素内容も一緒に移動させる。

【０１２２】省略タグ・属性補完部１５４は、タグ・属性リスト保持部１５２に保持されたタグ・属性リストに従って、タグ・属性並び替え部１５３による処理を施されたＸＭＬ文書中で省略されているタグや属性を補完するものである。つまり、省略タグ・属性補完部１５４は、入力されたＸＭＬ文書とタグリストとを比較して、そのＸＭＬ文書中から欠落しているタグや属性を検出すると、欠落タグや欠落属性に対応するタグあるいは属性名をタグ・属性リストから読み出し、その欠落タグや欠落属性を補完するものである。このとき、対になる開始

タグと終了タグとを補完する場合、省略タグ・属性補完部１５４は、これらの開始タグと終了タグとの間に記述されるべき要素内容を空のままとする。また、属性を補完する場合、省略タグ・属性補完部１５４は、属性値としてデフォルト値等を設定する。

【０１２３】そして、タグ・属性並び替え部１５３および省略タグ・属性補完部１５４による処理を施されたＸＭＬ文書は、第３実施形態の構造化文書圧縮装置１３０に入力されるようになっている。

【０１２４】次に、上述のごとく構成された、第４実施形態の変形例の構造化文書圧縮装置１５０の動作について説明する。図１２に示す構造化文書圧縮装置１５０においては、まず、圧縮対象のＸＭＬ文書を入力部１５１により取り込み、タグ・属性並び替え部１５３において、そのＸＭＬ文書とタグ・属性リストとが比較され、万一、ＸＭＬ文書中にタグや属性の記述順序の逆転等の不備がある場合には、ＸＭＬ文書におけるタグや属性の記述順序がタグ・属性リストにおけるタグや属性の配列順序に合うように、圧縮前のＸＭＬ文書のタグや属性が並び替えられる。

【０１２５】そして、並び替え処理を施されたＸＭＬ文書は、省略タグ・属性補完部１５４に入力され、この省略タグ・属性補完部１５４において、そのＸＭＬ文書とタグ・属性リストとが比較され、そのＸＭＬ文書中から欠落しているタグや属性が検出されると、欠落タグや欠落属性に対応するタグや属性がタグ・属性リストから読み出され、その欠落タグや欠落属性が補完される。

【０１２６】その結果、タグや属性の記述に不備のあったＸＭＬ文書が、タグ・属性リストに応じたデータ構造をもつＸＭＬ文書に修正され正規化（整頓）される。つまり、構造化文書圧縮装置１３０による圧縮対象である、全てのＸＭＬ文書に対し、タグ・属性並び替え部１５３および省略タグ・属性補完部１５４による処理を施すことによって、全てのＸＭＬ文書が、タグ・属性リストに応じたデータ構造をもつＸＭＬ文書となるように正規化される。

【０１２７】そして、上述のごとく正規化されたＸＭＬ文書が、構造化文書圧縮装置１３０に入力され、第３実施形態で前述したように圧縮される。なお、当然、不備のないＸＭＬ文書は、タグ・属性並び替え部１５３および省略タグ・属性補完部１５４をそのまま通過して、構造化文書圧縮装置１３０に入力される。

【０１２８】このように、本発明の第４実施形態における構造化文書圧縮装置１４０，１５０によれば、所定のデータ構造を定義する、タグリストまたはタグ・属性リストに従って、圧縮前のＸＭＬ文書のタグや属性が所定の順序に並び替えられるとともに、ＸＭＬ文書中で省略されているタグや属性が補完される。これにより、タグまたは属性の記述順序の逆転や、タグまたは属性の記述の欠落といった不備をもつＸＭＬ文書は、所定のデータ構造を有するように正規化される。

【０１２９】従って、同一のデータ構造を有する多数のＸＭＬ文書を圧縮処理対象とする場合、上述のような不備をもつＸＭＬ文書が含まれていても、圧縮処理前に、圧縮処理対象の全てのＸＭＬ文書が、タグリストもしくはタグ・属性リストで定義された所定のデータ構造を有するように正規化される。これにより、多数のＸＭＬ圧縮文書を、一つのタグリストまたはタグ・属性リストによって確実に管理し取り扱うことが可能になる。

【０１３０】〔５〕第５実施形態の説明

図１３は本発明の第５実施形態における構造化文書圧縮装置および構造化文書復元装置の構成を示すブロック図である。図１３に示す構造化文書圧縮装置１６０および構造化文書復元装置２６０は、ＸＭＬ文書に対する処理を行なうための構造化文書処理システムに含まれて、この構造化文書処理システムの一部を構成するものである。

【０１３１】この第５実施形態の構造化文書圧縮装置１６０は、ＸＭＬ文書を圧縮するためのもので、図１３に示すように、上述した構造化文書圧縮装置１２０，１３０，１４０，１５０のいずれか一つに対し、さらに、入力部１６１，サブ文書抽出部１６２，サブ文書統合部１６３および出力部１６４をそなえて構成されている。

【０１３２】また、第５実施形態の構造化文書復元装置２６０は、構造化文書圧縮装置１６０により生成されたＸＭＬ圧縮文書（圧縮データ）をＸＭＬ文書に復元するためのもので、図１３に示すように、上述した構造化文書復元装置２２０，２３０のいずれか一方に対し、さらに、入力部１６１，サブ文書抽出部１６２，サブ文書統合部１６３および出力部１６４をそなえて構成されている。

【０１３３】ここで、構造化文書圧縮装置１６０および構造化文書復元装置２６０は、同一のコンピュータ上、もしくは、それぞれ異なるコンピュータ上にそなえられている。そして、構造化文書圧縮装置１６０における各種機能は、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。同様に、構造化文書復元装置２６０における各種機能は、コンピュータ上で所定のプログラム（構造化文書復元プログラム）を実行することにより実現されるようになっている。

【０１３４】なお、図１３に示すように、構造化文書圧縮装置１６０と構造化文書復元部２６０とは、サブ文書抽出部１６２とサブ文書統合部１６３との間に配置される装置が異なる以外は、全く同じ構成になっている。さて、図１３に示す構造化文書圧縮装置１６０において、入力部１６１は、圧縮対象のＸＭＬ文書を、ハードディスク等（例えば図１の符号３００参照）から取り込むものであり、サブ文書抽出部１６２は、入力されたＸＭＬ文書から、所定の要素名をもつ開始タグと終了タグとで

囲まれた領域をサブ文書として抽出し、そのサブ文書を構造化文書圧縮装置１２０，１３０，１４０，１５０のいずれか一つ（以下、第５実施形態では、構造化文書圧縮装置１２０とする）に出力するものである。

【０１３５】そして、サブ文書統合部１６３は、構造化文書圧縮装置１２０からサブ文書の圧縮結果を受け、その圧縮結果とサブ文書以外の部分とを統合するものであり、出力部１６４は、サブ文書統合部１６３によって統合されたＸＭＬ圧縮文書を、圧縮結果として、ハードディスク等（例えば図１，図２，図１５～図１７の符号４００，４１０，４２０，４４０参照）に出力・格納するものである。

【０１３６】一方、図１３に示す構造化文書復元装置２６０において、入力部１６１は、復元対象のＸＭＬ圧縮文書を、記憶媒体等（例えば図１，図２，図１５～図１７に示すハードディスク４００，４１０，４２０，４４０）から取り込むものであり、サブ文書抽出部１６２は、入力されたＸＭＬ圧縮文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出し、そのサブ文書を構造化文書復元装置２２０もしくは２３０（以下、第５実施形態では、構造化文書復元装置２２０とする）に出力するものである。

【０１３７】そして、サブ文書統合部１６３は、構造化文書復元装置２２０からサブ文書の復元結果を受け、その復元結果とサブ文書以外の部分とを統合するものであり、出力部１６４は、サブ文書統合部１６３によって統合されたＸＭＬ文書を、復元結果として、記憶媒体等（例えば図１に示すハードディスク３００）に出力・格納するものである。

【０１３８】なお、構造化文書復元装置２６０において構造化文書復元装置２２０を用いる場合、その構造化文書復元装置２２０において用いられるタグリストは、サブ文書におけるタグを出現順序に従ってリスト化して予め生成されたもので、データベース（図示略）等から取得される。また、構造化文書復元装置２６０において構造化文書復元装置２３０を用いる場合、その構造化文書復元装置２３０において用いられるタグ・属性リストは、サブ文書におけるタグや属性を出現順序に従ってリスト化して予め生成されたもので、やはり、データベース（図示略）等から取得される。

【０１３９】次に、上述のごとく構成された、第５実施形態の構造化文書圧縮装置１６０および構造化文書復元装置２６０の動作について説明する。図１３に示す構造化文書圧縮装置１６０においては、まず、圧縮対象のＸＭＬ文書を入力部１６１により取り込み、サブ文書抽出部１６２により、そのＸＭＬ文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域がサブ文書として抽出され、そのサブ文書が構造化文書圧縮装置１２０に出力される。

【０１４０】そして、構造化文書圧縮装置１２０においては、入力されたサブ文書に対し、第２実施形態で前述した圧縮処理が施され、タグを所定の区切りコードに置き換えるようにして圧縮されたサブ文書が生成される。圧縮されたサブ文書は、サブ文書統合部１６３によりサブ文書以外の部分と統合され、統合されたＸＭＬ文書が圧縮結果として出力部１６４から出力される。

【０１４１】ここで、図１４（Ａ）および図１４（Ｂ）を参照しながら、第５実施形態における具体的なＸＭＬ文書の圧縮状態について説明する。なお、図１４（Ａ）および図１４（Ｂ）はいずれも第５実施形態におけるデータ例を示すもので、図１４（Ａ）は複数のサブ文書を含むＸＭＬ文書の一例を示す図、図１４（Ｂ）は図１４（Ａ）に示すＸＭＬ文書の圧縮状態を示す図である。

【０１４２】図１４（Ａ）には、圧縮前つまり圧縮対象のＸＭＬ文書の一例として、同一のデータ構造をもつ複数（図中３つ）のサブ文書を含む、発注伝票についてのＸＭＬ文書が示されている。この図１４（Ａ）に示すＸＭＬ文書では、開始タグ＜商品＞と終了タグ＜/商品＞とで囲まれた領域が３つ存在し、これらの領域は全く同じデータ構造を有している。つまり、各領域においては、メーカ，製品番号および価格についてのタグと要素内容とが記述されている。ただし、これらの領域に記述された要素内容は異なっている。

【０１４３】図１４（Ａ）に示すＸＭＬ文書を、構造化文書圧縮装置１６０により圧縮する場合、サブ文書抽出部１６２において、サブ文書の抽出基準として開始タグ＜商品＞および終了タグ＜/商品＞を予め設定しておくことにより、図１４（Ａ）に示すＸＭＬ文書から、開始タグ＜商品＞と終了タグ＜/商品＞とにより囲まれた、３つの領域がサブ文書として抽出される。

【０１４４】抽出された各サブ文書に対し、構造化文書圧縮装置１２０による圧縮処理を施した結果、図１４（Ａ）に示すサブ文書中のタグは区切りコード「，」に置き換えられる。そして、置換処理後のサブ文書とサブ文書以外の部分とをサブ文書統合部１６３により統合すると、図１４（Ｂ）に示すようなＸＭＬ圧縮文書が生成される。

【０１４５】一方、図１３に示す構造化文書復元装置２６０においては、まず、例えば図１４（Ｂ）に示すようなＸＭＬ圧縮文書を復元対象として入力部１６１により取り込み、圧縮処理時と同様、サブ文書抽出部１６２により、そのＸＭＬ圧縮文書から、開始タグ＜商品＞と終了タグ＜/商品＞とで囲まれた領域（実質的なＸＭＬ圧縮文書）がサブ文書として抽出され、そのサブ文書（ＸＭＬ圧縮文書）が構造化文書復元装置２２０に出力される。

【０１４６】そして、構造化文書復元装置２２０においては、入力されたサブ文書に対し、第２実施形態で前述した復元処理が施されて、区切りコード「，」が適切なタグに復元され、サブ文書が元のＸＭＬ文書に復元され

る。復元されたサブ文書は、サブ文書統合部１６３によりサブ文書以外の部分と統合され、統合されたＸＭＬ文書が復元結果として出力部１６４から出力される。

【０１４７】このように、本発明の第５実施形態における構造化文書圧縮装置１６０によれば、一つのＸＭＬ文書中に、同一のデータ構造を有する領域（サブ文書）が複数存在する場合、ＸＭＬ文書からそのサブ文書が抽出され、各サブ文書中のタグを区切りコード「，」に置き換えることにより、各サブ文書について、ＸＭＬ文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、ＸＭＬ文書が圧縮されてＸＭＬ文書のデータ量を削減することができる。

【０１４８】従って、第１実施形態や第２実施形態と同様、ＸＭＬ文書を格納するための記憶領域（例えば図１，図２，図１５～図１７に示すハードディスク４００，４１０，４２０，４４０）の容量を削減することができるとともに、ＸＭＬ文書データの伝送速度を高速化することができる。

【０１４９】また、第５実施形態の構造化文書復元装置２６０によれば、上述のような圧縮を施されたサブ文書を含むＸＭＬ文書が復元対象になると、その復元対象のサブ文書中で検出された区切りコード「，」を、サブ文書についてのタグリスト中のタグと対応させながら、所定のタグに置き換えるという簡易な置換処理により、復元対象の書を極めて容易に元の構造化文書に復元することができる。

【０１５０】〔６〕タグリストの一括管理手法の説明

次に、複数種類のデータ構造（つまり複数種類のタグリスト）を一つの構造化文書処理システムで管理する場合の、本実施形態におけるタグリストの一括管理手法について、図１５～図１８を参照しながら説明する。なお、図１５～図１７は、それぞれ、本実施形態におけるタグリストの一括管理手法の第１例～第３例を説明するための図、図１８本実施形態におけるタグリストの一括管理手法の第２例および第３例におけるタグリスト識別情報の付加例を示す図である。である。

【０１５１】図１５に示す構造化文書処理システムでは、ハードディスク（データベース）４１０において、構造化文書圧縮装置１２０～１６０により生成された複数（図１５では３つ）のＸＭＬ圧縮文書が格納される。ここで、３つのＸＭＬ圧縮文書には、それぞれ、識別情報（識別子）１～３が付与されているものとする。

【０１５２】そして、ハードディスク４１０には、タグリスト群保持部４１１およびタグリスト管理部４１２が保持されている。タグリスト群保持部４１１は、処理対象となるＸＭＬ文書のデータ構造に対応した複数種類（図１５では２種類）のタグリストを予め保持するものである。ここで、２種類のタグリストには、それぞれ、タグリスト識別情報（タグリスト識別子）Ａ，Ｂが付与されているものとする。

【０１５３】タグリスト管理部４１２は、構造化文書圧縮装置１２０～１６０によって生成されたＸＭＬ圧縮文書の識別情報１～３と、タグリスト群保持部４１１に保持されているタグリストＡ，Ｂとの対応関係をテーブルによって一括管理するものである。例えば図１５に示すタグリスト管理部４１２のテーブルによれば、ＸＭＬ圧縮文書１，２，３とタグリストＡ，Ａ，Ｂとがそれぞれ対応関係にある。このタグリスト管理部４１２により、ハードディスク４１０に保持されている複数のＸＭＬ圧縮文書をそれぞれ復元する際に必要なタグリストが、一括管理される。

【０１５４】従って、構造化文書復元装置２２０，２３０，２６０においてＸＭＬ圧縮文書を復元する際には、そのＸＭＬ圧縮文書の識別情報をキーにしてタグリスト管理部４１２のテーブルを検索することにより、そのＸＭＬ圧縮文書の識別情報に対応した、タグリスト識別情報を得る。そして、構造化文書復元装置２１０～２３０，２６０は、そのタグリスト識別情報により特定されるタグリストを、ハードディスク４１０のタグリスト群保持部４１１から読み出し、上述したようなＸＭＬ圧縮文書の復元処理に使用する。

【０１５５】図１６に示す構造化文書処理システムでは、ハードディスク（データベース）４２０において、構造化文書圧縮装置１１０～１５０により生成された複数（図１６では３つ）のＸＭＬ圧縮文書１～３が格納されるとともに、図１５に示したものと同様のタグリスト群保持部４１１が保持されている。

【０１５６】また、ハードディスク４２０には、構造化文書圧縮装置１２０～１６０がアクセス可能に接続されるとともに構造化文書復元装置２２０，２３０，２６０がアクセス可能に接続されており、構造化文書圧縮装置１２０～１６０には、タグリスト識別情報付加部１７１がそなえられるとともに、構造化文書復元装置２２０，２３０，２６０には、タグリスト識別情報取得部１７２がそなえられている。

【０１５７】タグリスト識別情報付加部１７１は、構造化文書圧縮装置１２０～１６０によって生成されたＸＭＬ圧縮文書に、そのＸＭＬ圧縮文書に対応するタグリストを特定するためのタグリスト識別情報を付加するものであり、タグリスト識別情報取得部１７２は、ＸＭＬ圧縮文書に付加されたタグリスト識別情報を取得するものである。

【０１５８】従って、構造化文書圧縮装置１２０～１６０においてＸＭＬ圧縮文書が生成されると、そのＸＭＬ圧縮文書に対応するタグリスト識別情報（識別子）を、例えば図１８に示すごとく、タグリスト識別情報付加部１７１によりＸＭＬ圧縮文書の開始タグ内に属性として書き込んで付加する。なお、図１６に示す例では、ＸＭＬ圧縮文書１～３のそれぞれにタグリスト識別情報Ａ，Ａ，Ｂが付加されている。また、図１８では、ＸＭＬ圧

縮文書１または２における開始タグ＜商品＞の中に、タグリスト識別情報Ａが属性「tag=‘タグリストＡ’」として記入された例が示されている。

【０１５９】一方、構造化文書復元装置２２０，２３０，２６０においてＸＭＬ圧縮文書を復元する際には、そのＸＭＬ圧縮文書に付加されているタグリスト識別情報を、タグリスト識別情報取得部１７２により取得する。そして、構造化文書復元装置２２０，２３０，２６０は、そのタグリスト識別情報により特定されるタグリストを、ハードディスク４２０のタグリスト群保持部４１１から読み出し、上述したようなＸＭＬ圧縮文書の復元処理に使用する。

【０１６０】図１７に示す構造化文書処理システムでは、管理サーバ６００におけるハードディスク（データベース）４３０に、図１５に示したものと同様のタグリスト群保持部４１１が保持されている。また、管理サーバ６００は、ＬＡＮ等のネットワーク７００を介して構造化文書圧縮装置１２０～１６０や構造化文書復元装置２２０，２３０，２６０と通信可能に接続されるほか、これらの構造化文書圧縮装置１２０～１６０や構造化文書復元装置２２０，２３０，２６０は、ハードディスク（データベース）４４０にアクセス可能に接続されている。このハードディスク４４０には、図１６に示した例と同様、構造化文書圧縮装置１２０～１６０において生成されそれぞれタグリスト識別情報を付加されたＸＭＬ圧縮文書が格納されている。

【０１６１】従って、構造化文書圧縮装置１２０～１６０においてＸＭＬ圧縮文書が生成されると、図１６に示したシステムと同様、そのＸＭＬ圧縮文書に対応するタグリスト識別情報（識別子）を、例えば図１８に示すごとく、タグリスト識別情報付加部１７１によりＸＭＬ圧縮文書の開始タグ内に属性として書き込んで付加する。なお、図１７に示す例でも、ＸＭＬ圧縮文書１～３のそれぞれにはタグリスト識別情報Ａ，Ａ，Ｂが付加されている。

【０１６２】一方、構造化文書復元装置２２０，２３０，２６０においてＸＭＬ圧縮文書を復元する際には、そのＸＭＬ圧縮文書に付加されているタグリスト識別情報を、タグリスト識別情報取得部１７２により取得する。そして、構造化文書復元装置２２０，２３０，２６０は、そのタグリスト識別情報により特定されるタグリストを、ハードディスク４３０（即ち、管理サーバ６００上）のタグリスト群保持部４１１から、ネットワーク７００経由で読み出し、上述したようなＸＭＬ圧縮文書の復元処理に使用する。

【０１６３】このように、図１５～図１７に示す構造化文書処理システムによれば、ＸＭＬ圧縮文書とタグリストとの対応関係をタグリスト管理部４１２によって管理したり、ＸＭＬ圧縮文書に対応するタグリストを特定するためのタグリスト識別情報をＸＭＬ圧縮文書に付加したりすることで、タグリスト群が一括管理され、ＸＭＬ圧縮文書とタグリストとの対応関係を確実に把握することができ、ＸＭＬ圧縮文書を、そのＸＭＬ圧縮文書に対応したタグリストに基づいて復元することができる。

【０１６４】従って、ＸＭＬ文書に対し圧縮・復元処理を施しながらＸＭＬ文書を取り扱うシステムにおいて、異なる種類のデータ構造（即ちタグリスト）をもつＸＭＬ文書（ＸＭＬ圧縮文書）が混在しても、混乱を招くことなく、各ＸＭＬ圧縮文書に応じたタグリストを確実に取得して復元処理を行なうことができる。

【０１６５】また、図１７に示す構造化文書処理システムによれば、タグリスト群を管理サーバ６００上で保持・管理し、ネットワーク７００を介して処理に必要なタグリストを管理サーバ６００から読み出すように構成することにより、タグリスト群が一括管理される。従って、複数種類のタグリストを構造化文書圧縮装置毎や構造化文書復元装置毎に管理する必要がなくなり、複数の構造化文書圧縮装置や構造化文書復元装置によって共用することができる。

【０１６６】なお、図１５～図１８に示したシステムでは、タグリスト群を一括管理する場合について説明したが、タグ・属性リスト群についても上述と同様にして一括管理することができる。

【０１６７】〔７〕その他

なお、本発明は上述した実施形態に限定されるものではなく、本発明の趣旨を逸脱しない範囲で種々変形して実施することができる。例えば、上述した実施形態では、構造化文書がＸＭＬである場合について説明したが、本発明は、これに限定されるものではなく、タグを用いて記述される、ＸＭＬと同様の構造化文書（ＳＧＭＬ等）に対し、上述した実施形態と同様に適用され、上述と同様の作用効果を得ることができる。

【０１６８】また、上述した実施形態では、区切りコードとして「，」や「/」や「=」を用いた場合について説明したが、本発明は、これに限定されるものではなく、要素内容の記述に使用されることのない、他の文字あるいは記号を区切りコードとして用いてもよく、この場合も、上述した実施形態と同様の作用効果を得ることができる。

【０１６９】〔８〕付記

（付記１）　同一のデータ構造を有する複数の構造化文書を圧縮する装置であって、該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力

する出力部とをそなえたことを特徴とする、構造化文書圧縮装置。

【０１７０】（付記２）　該構造化文書圧縮部が、前記の各構造化文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを前記所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえて構成されていることを特徴とする、付記１記載の構造化文書圧縮装置。

【０１７１】（付記３）　構造化文書を圧縮する装置であって、該構造化文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえたことを特徴とする、構造化文書圧縮装置。

【０１７２】（付記４）　構造化文書を圧縮する装置であって、該構造化文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出するサブ文書抽出部と、該サブ文書抽出部により抽出された該サブ文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえたことを特徴とする、構造化文書圧縮装置。

【０１７３】（付記５）　該タグ検出部により検出された該タグが属性値をもつ属性付きタグであるか否かを検出する属性付きタグ検出部と、該属性付きタグ検出部により検出された該属性付きタグを前記属性値および所定の区切りコードに置き換えて圧縮する属性タグ付きタグ圧縮部とをそなえたことを特徴とする、付記３または付記４に記載の構造化文書圧縮装置。

【０１７４】（付記６）　所定のデータ構造を定義すべく所定の順序でタグを並べたタグリストを予め保持するタグリスト保持部と、該タグリスト保持部に保持された前記タグリストに従って、圧縮前の前記構造化文書のタグを前記所定の順序に並び替えるタグ並び替え部と、該タグリスト保持部に保持された前記タグリストに従って、該構造化文書中で省略されているタグを補完する省略タグ補完部とをそなえたことを特徴とする、付記３または付記４に記載の構造化文書圧縮装置。

【０１７５】（付記７）　所定のデータ構造を定義すべく所定の順序で並べたタグと属性名とをもつタグ・属性リストを予め保持するタグ・属性リスト保持部と、該タグ・属性リスト保持部に保持された前記タグ・属性リストに従って、圧縮前の前記構造化文書のタグおよび属性を前記所定の順序に並び替えるタグ・属性並び替え部と、該タグ・属性リスト保持部に保持された前記タグ・属性リストに従って、該構造化文書中で省略されているタグおよび属性を補完する省略タグ・属性補完部とをそなえたことを特徴とする、付記５記載の構造化文書圧縮装置。

【０１７６】（付記８）　同一のデータ構造を有する複数の構造化文書を圧縮する方法であって、該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得し、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成し、前記一つのタグリストと、該複数の構造化文書のそれぞれについて生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力することを特徴とする、構造化文書圧縮方法。

【０１７７】（付記９）　構造化文書を圧縮する方法であって、該構造化文書中のタグを検出し、検出された該タグを所定の区切りコードに置き換えて圧縮することを特徴とする、構造化文書圧縮方法。

【０１７８】（付記１０）　構造化文書を圧縮する方法であって、該構造化文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出し、該サブ文書中のタグを検出し、検出された該タグを所定の区切りコードに置き換えて圧縮することを特徴とする、構造化文書圧縮方法。

【０１７９】（付記１１）　同一のデータ構造を有する複数の構造化文書を圧縮する機能をコンピュータにより実現するための構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書圧縮プログラムが、該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部、および、該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部として、該コンピュータに機能させることを特徴とする、構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体。

【０１８０】（付記１２）　構造化文書を圧縮する機能をコンピュータにより実現するための構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書圧縮プログラムが、該構造化文書中のタグを検出するタグ検出部、および、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部として、該コンピュータを機能させることを特徴とする、構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体。

【０１８１】（付記１３）　構造化文書を圧縮する機能をコンピュータにより実現するための構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書圧縮プログラムが、該構造化文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出するサブ文書抽出部、該サブ文書抽出部により抽出された該サブ文書中のタグを

検出するタグ検出部、および、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部として、該コンピュータを機能させることを特徴とする、構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体。

【０１８２】（付記１４）　同一のデータ構造を有する複数の構造化文書中のタグを所定の区切りコードに置き換えることにより生成された複数の圧縮文書を、該複数の構造化文書におけるタグを出現順序に従ってリスト化したタグリストに基づいて復元する装置であって、該タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製する複製部と、該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえたことを特徴とする、構造化文書復元装置。

【０１８３】（付記１５）　構造化文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する装置であって、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書中の前記所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえたことを特徴とする、構造化文書復元装置。

【０１８４】（付記１６）　構造化文書において所定の要素名をもつ開始タグと終了タグとで囲まれた領域であるサブ文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する装置であって、該サブ文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書から該サブ文書を抽出するサブ文書抽出部と、該サブ文書抽出部により抽出された該サブ文書中の前記所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえたことを特徴とする、構造化文書復元装置。

【０１８５】（付記１７）　該圧縮文書中において、属性付きタグ内の属性が属性値および所定の区切りコードに置き換えられて圧縮されている場合、該圧縮文書における属性名を出現順序に従ってリスト化した属性リストを予め保持する属性リスト保持部と、該タグ復元部で復元対象となったタグが属性付きタグに復元されるべきものであるか否かを検出する属性付きタグ検出部と、該属性付きタグについての属性値と該属性リストにおける属性名とを対応させて、該属性付きタグ検出部により検出された該属性付きタグ内の該属性を復元する属性付きタグ復元部とをそなえたことを特徴とする、付記１５または付記１６に記載の構造化文書復元装置。

【０１８６】（付記１８）　同一のデータ構造を有する複数の構造化文書中のタグを所定の区切りコードに置き換えることにより生成された複数の圧縮文書を、該複数の構造化文書におけるタグを出現順序に従ってリスト化したタグリストに基づいて復元する方法であって、該タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製し、該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出すことを特徴とする、構造化文書復元方法。

【０１８７】（付記１９）　構造化文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する方法であって、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持し、該圧縮文書中の前記所定の区切りコードを検出し、検出された前記所定の区切りコードの位置と該タグリストにおけるタグの位置とを対応させながら、検出された前記所定の区切りコードを該タグリストにおける対応するタグに置き換えて復元することを特徴とする、構造化文書復元方法。

【０１８８】（付記２０）　構造化文書において所定の要素名をもつ開始タグと終了タグとで囲まれた領域であるサブ文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する方法であって、該サブ文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持し、該圧縮文書から該サブ文書を抽出し、抽出された該サブ文書中の前記所定の区切りコードを検出し、検出された前記所定の区切りコードの位置と該タグリストにおけるタグの位置とを対応させながら、検出された前記所定の区切りコードを該タグリストにおける対応するタグに置き換えて復元することを特徴とする、構造化文書復元方法。

【０１８９】（付記２１）　同一のデータ構造を有する複数の構造化文書中のタグを所定の区切りコードに置き換えることにより生成された複数の圧縮文書を、該複数の構造化文書におけるタグを出現順序に従ってリスト化したタグリストに基づいて復元する機能をコンピュータにより実現するための構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書復元プログラムが、該タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製する複製部、および、該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを

対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部として、該コンピュータを機能させることを特徴とする、構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体。

【０１９０】（付記２２） 構造化文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する機能をコンピュータにより実現するための構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書復元プログラムが、該圧縮文書中の前記所定の区切りコードを検出する区切りコード検出部、および、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストにおけるタグの位置と、該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部として、該コンピュータを機能させることを特徴とする、構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体。

【０１９１】（付記２３） 構造化文書において所定の要素名をもつ開始タグと終了タグとで囲まれた領域であるサブ文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する機能をコンピュータにより実現するための構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書復元プログラムが、該圧縮文書から該サブ文書を抽出するサブ文書抽出部、該サブ文書抽出部により抽出された該サブ文書中の前記所定の区切りコードを検出する区切りコード検出部、および、該サブ文書におけるタグを出現順序に従ってリスト化したタグリストにおけるタグの位置と、該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部として、該コンピュータを機能させることを特徴とする、構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体。

【０１９２】（付記２４） 同一のデータ構造を有する複数の構造化文書に対する処理を行なうべく、該複数の構造化文書を圧縮する構造化文書圧縮装置と、該構造化文書圧縮装置による圧縮データを該複数の構造化文書に復元する構造化文書復元装置とを含んで構成される構造化文書処理システムにおいて、該構造化文書圧縮装置が、該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部とをそなえて構成されるとともに、該構造化文書復元装置が、該複数の圧縮文書の復元結果を格納するメモリと、該タグリストに対応するデータ構造を該メモリ上に複製データ構造として展開・複製する複製部と、該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえて構成されたことを特徴とする、構造化文書処理システム。

【０１９３】（付記２５） 構造化文書に対する処理を行なうべく、該構造化文書を圧縮する構造化文書圧縮装置と、該構造化文書圧縮装置による圧縮データを該構造化文書に復元する構造化文書復元装置とを含んで構成される構造化文書処理システムにおいて、該構造化文書圧縮装置が、該構造化文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえて構成されるとともに、該構造化文書復元装置が、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書中の前記所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえて構成されたことを特徴とする、構造化文書処理システム。

【０１９４】（付記２６） 構造化文書に対する処理を行なうべく、該構造化文書を圧縮する構造化文書圧縮装置と、該構造化文書圧縮装置による圧縮データを該構造化文書に復元する構造化文書復元装置とを含んで構成される構造化文書処理システムにおいて、該構造化文書圧縮装置が、該構造化文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出するサブ文書抽出部と、該サブ文書抽出部により抽出された該サブ文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえて構成されるとともに、構造化文書復元装置が、該サブ文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書から該サブ文書を抽出するサブ文書抽出部と、該サブ文書抽出部により抽出された該サブ文書中の前記所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該

区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえて構成されたことを特徴とする、構造化文書処理システム。

【０１９５】（付記２７）　処理対象となりうる構造化文書のデータ構造に対応した複数のタグリストを予め保持するタグリスト群保持部と、該構造化文書圧縮装置によって生成された該圧縮文書と、該タグリスト群保持部に保持されている該タグリストとの対応関係を管理するタグリスト管理部とをそなえたことを特徴とする、付記２５または付記２６に記載の構造化文書処理システム。

【０１９６】（付記２８）　処理対象となりうる構造化文書のデータ構造に対応した複数のタグリストを予め保持するタグリスト群保持部と、該構造化文書圧縮装置によって生成された該圧縮文書に、該圧縮文書に対応するタグリストを特定するためのタグリスト識別情報を付加するタグリスト識別情報付加部と、該圧縮文書に付加された前記タグリスト識別情報を取得するタグリスト識別情報取得部とをそなえ、該構造化文書復元装置が、該タグリスト識別情報取得部によって取得された前記タグリスト識別情報に対応する該タグリストを用いて、該圧縮文書を復元することを特徴とする、付記２５または付記２６に記載の構造化文書処理システム。

【０１９７】（付記２９）　該タグリスト群保持部が管理サーバ上に配置され、処理に必要なタグリストが、ネットワークを介して該管理サーバ上の該タグリスト群保持部から読み出されることを特徴とする、付記２７または付記２８に記載の構造化文書処理システム。

【０１９８】

【発明の効果】以上詳述したように、本発明の構造化文書圧縮装置（請求項１，２）および構造化文書復元装置（請求項３，４）並びに構造化文書処理システム（請求項５）によれば、以下のような効果ないし利点を得ることができる。

（１）本発明により生成される圧縮文書では、タグが所定の区切りコードに置換されているだけで、データ内容（要素内容）はそのまま記述されているので、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性を生かしたまま、構造化文書を圧縮して構造化文書のデータ量を削減することができる。従って、構造化文書を格納するための記憶領域の容量を削減することができるとともに構造化文書データの伝送速度を高速化することができる（請求項１，５）。

【０１９９】（２）複数の構造化文書の圧縮結果は、データ構造（一つのタグリスト）とデータ内容（複数の圧縮文書）とに分離されているので、一つのタグリストに対する解析処理を一度だけ行なって、複数の圧縮文書に共通のデータ構造を取得してしまえば、後は、取得されたデータ構造を複製して流用することにより、圧縮文書毎に一々タグ解析を行なう必要がなくなる。従って、同一のデータ構造を有する多数の構造化文書を取り扱う際に、無駄なタグ解析を行なう必要が一切なくなり、タグ解析の負荷が大幅に低減され、構造化文書をメモリに展開する際の処理速度を飛躍的に高速化することができる（請求項１，３，５）。

【０２００】（３）構造化文書中で検出されたタグを所定の区切りコードに置換するという極めて単純な置換処理によって、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を圧縮して構造化文書のデータ量を削減することができる。従って、構造化文書を格納するための記憶領域の容量を削減することができるとともに構造化文書データの伝送速度を高速化することができる（請求項２）。このような圧縮を行なった場合、圧縮文書中で検出された所定の区切りコードを、その圧縮文書についてのタグリスト中のタグと対応させながら、所定のタグに置き換えるという簡易な置換処理によって、圧縮文書を極めて容易に元の構造化文書に復元することができる（請求項４）。

【０２０１】（４）一つの構造化文書中に、同一のデータ構造を有する領域（サブ文書）が複数存在する場合、構造化文書から、そのサブ文書が、所定の要素名をもつ開始タグと終了タグとで囲まれた領域として抽出され、各サブ文書中のタグを所定の区切りコードに置き換えることにより、各サブ文書について、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を圧縮して構造化文書のデータ量を削減することができる。従って、構造化文書を格納するための記憶領域の容量を削減することができるとともに構造化文書データの伝送速度を高速化することができる。このような圧縮を行なった場合、復元対象におけるサブ文書中で検出された所定の区切りコードを、サブ文書についてのタグリスト中のタグと対応させながら、所定のタグに置き換えるという簡易な置換処理によって、復元対象の文書を極めて容易に元の構造化文書に復元することができる。

【０２０２】（５）タグが属性値をもつ属性付きタグである場合には、その属性付きタグを属性値および所定の区切りコードに置き換えて圧縮する。これにより、圧縮文書において属性値がそのまま記述されるので、属性値の視認性を保ちながら構造化文書の圧縮を行なうことができる。このような圧縮を施された属性付きタグが復元対象になると、その属性付きタグについての属性値とその圧縮文書についての属性リスト中の属性名とを対応させることにより、属性付きタグを極めて容易に復元することができる。

【０２０３】（６）所定のデータ構造を定義する、タグリストまたはタグ・属性リストに従って、圧縮前の前記構造化文書のタグまたは属性を所定の順序に並び替えるとともに、構造化文書中で省略されているタグまたは属

性を補完することにより、タグまたは属性の記述順序の逆転や、タグまたは属性の記述の欠落といった不備をもつ構造化文書は、所定のデータ構造を有するように正規化される。従って、同一のデータ構造を有する多数の構造化文書を圧縮処理対象とする場合、上述のような不備をもつ構造化文書が含まれていても、圧縮処理前に、圧縮処理対象の全ての構造化文書が、タグリストまたはタグ・属性リストで定義された所定のデータ構造を有するように正規化される。これにより、多数の構造化文書（圧縮文書）を、一つのタグリストまたはタグ・属性リストによって確実に管理し取り扱うことが可能になる。

【０２０４】（７）圧縮文書とタグリストとの対応関係をタグリスト管理部によって管理したり、圧縮文書に対応するタグリストを特定するためのタグリスト識別情報を圧縮文書に付加したりすることで、タグリスト群が一括され、圧縮文書とタグリストとの対応関係を確実に把握でき、圧縮文書を、その圧縮文書に対応したタグリストに基づいて復元することができる。従って、構造化文書に対し圧縮・復元処理を施しながら構造化文書を取り扱うシステムにおいて、異なる種類のデータ構造（即ちタグリスト）をもつ構造化文書（圧縮文書）が混在しても、混乱を招くことなく、各圧縮文書に応じたタグリストを確実に取得して復元処理を行なうことができる。

【０２０５】（８）タグリスト群を管理サーバ上で保持・管理し、ネットワークを介して処理に必要なタグリストを管理サーバから読み出すように構成することにより、タグリスト群が一括管理される。従って、複数種類のタグリストを圧縮装置毎や復元装置毎に管理する必要がなくなり、複数の圧縮装置や復元装置によって共用することができる。

【図面の簡単な説明】

【図１】本発明の第１実施形態における構造化文書圧縮装置の構成を示すブロック図である。

【図２】本発明の第１実施形態における構造化文書復元装置（メモリ展開部）の構成を示すブロック図である。

【図３】（A）～（C）はいずれも第１実施形態におけるデータ例を示すもので、（A）はＸＭＬ文書の一例を示す図、（Ｂ）は（Ａ）に示すＸＭＬ文書から得られたタグリストを示す図、（Ｃ）は（Ａ）に示すＸＭＬ文書の圧縮状態を示す図である。

【図４】本発明の第２実施形態における構造化文書圧縮装置の構成を示すブロック図である。

【図５】本発明の第２実施形態における構造化文書復元装置の構成を示すブロック図である。

【図６】（A）～（Ｄ）はいずれも第２実施形態におけるデータ例を示すもので、（A）はＸＭＬ文書の一例を示す図、（Ｂ）は（Ａ）に示すＸＭＬ文書に対応するタグリストを示す図、（Ｃ）は（Ａ）に示すＸＭＬ文書の圧縮状態の一例を示す図、（Ｄ）は（Ａ）に示すＸＭＬ文書の圧縮状態の他例を示す図である。

【図７】本発明の第３実施形態における構造化文書圧縮装置の構成を示すブロック図である。

【図８】本発明の第３実施形態における構造化文書復元装置の構成を示すブロック図である。

【図９】（A）～（Ｄ）はいずれも第３実施形態におけるデータ例を示すもので、（A）はＸＭＬ文書の一例を示す図、（Ｂ）は（Ａ）に示すＸＭＬ文書に対応するタグ・属性リストを示す図、（Ｃ）は（Ａ）に示すＸＭＬ文書の圧縮状態の一例を示す図、（Ｄ）は（Ａ）に示すＸＭＬ文書の圧縮状態の他例を示す図である。

【図１０】本発明の第４実施形態における構造化文書圧縮装置の要部構成を示すブロック図である。

【図１１】（A）～（C）はいずれも第４実施形態におけるデータ例を示すもので、（A）はタグリストの一例を示す図、（Ｂ）はタグの記述に不備のあるＸＭＬ文書の一例を示す図、（Ｃ）は（Ｂ）に示すＸＭＬ文書を（Ａ）に示すタグリストに従って正規化した結果を示す図である。

【図１２】本発明の第４実施形態における構造化文書圧縮装置の変形例の要部構成を示すブロック図である。

【図１３】本発明の第５実施形態における構造化文書圧縮装置および構造化文書復元装置の構成を示すブロック図である。

【図１４】（A）および（Ｂ）はいずれも第５実施形態におけるデータ例を示すもので、（A）は複数のサブ文書を含むＸＭＬ文書の一例を示す図、（Ｂ）は（Ａ）に示すＸＭＬ文書の圧縮状態を示す図である。

【図１５】本実施形態におけるタグリストの一括管理手法の第１例を説明するための図である。

【図１６】本実施形態におけるタグリストの一括管理手法の第２例を説明するための図である。

【図１７】本実施形態におけるタグリストの一括管理手法の第３例を説明するための図である。

【図１８】本実施形態におけるタグリストの一括管理手法の第２例および第３例におけるタグリスト識別情報の付加例を示す図である。

【図１９】一般的なユニバーサルデータ圧縮について説明するための図である。

【図２０】ＸＭＬ文書を取り扱う一般的なシステムの構成例を示すブロック図である。

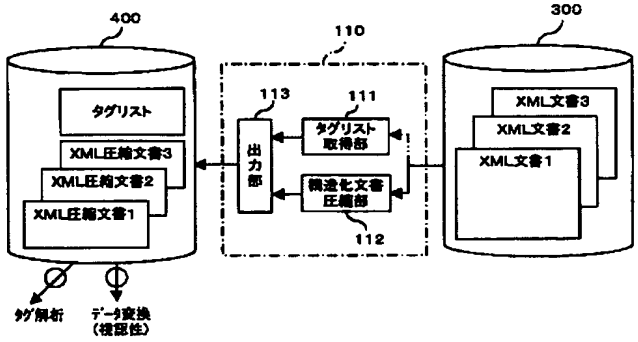
【符号の説明】

１１０　構造化文書圧縮装置
１１１　タグリスト取得部
１１２　構造化文書圧縮部
１１３　出力部
１２０　構造化文書圧縮装置
１２１　入力部
１２２　タグ検出部
１２３　タグ圧縮部
１２４　出力部

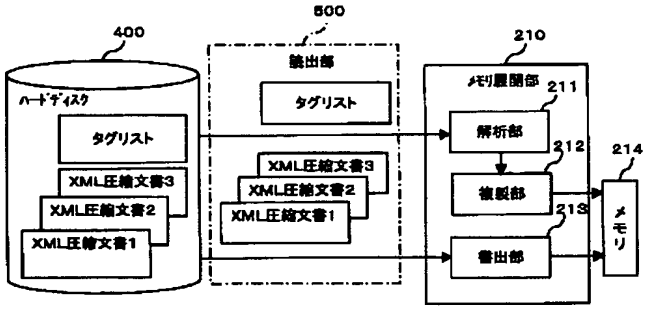
１３０　構造化文書圧縮装置
１３１　属性付きタグ検出部
１３２　属性付きタグ圧縮部
１４０　構造化文書圧縮装置
１４１　入力部
１４２　タグリスト保持部
１４３　タグ並び替え部
１４４　省略タグ補完部
１５０　構造化文書圧縮装置
１５１　入力部
１５２　タグ・属性リスト保持部
１５３　タグ・属性並び替え部
１５４　省略タグ・属性補完部
１６０　構造化文書圧縮装置
１６１　入力部
１６２　サブ文書抽出部
１６３　サブ文書統合部
１６４　出力部
１７１　タグリスト識別情報付加部
１７２　タグリスト識別情報取得部
２１０　メモリ展開部（構造化文書復元装置）
２１１　解析部

２１２　複製部
２１３　書出部
２１４　メモリ
２２０　構造化文書復元装置
２２１　入力部
２２２　タグリスト保持部
２２３　区切りコード検出部
２２４　タグ復元部
２２５　出力部
２３０　構造化文書復元装置
２３１　属性リスト保持部
２３２　属性付きタグ検出部
２３３　属性付きタグ復元部
２６０　構造化文書復元装置
３００，４００，４１０，４２０，４３０，４４０　ハードディスク（データベース）
４１１　タグリスト群保持部
４１２　タグリスト管理部
５００　読出部
６００　管理サーバ
７００　ネットワーク

【図１】



【図２】



【図５】

【図３】

(A)

```
<発注伝票>
<発注者><名前>STUV</名前>
        <電話番号>1111</電話番号>
</発注者>
<商品><メーカ>A社</メーカ>
        <製品番号>1234</製品番号>
        <製品名>ABCD</製品名>
        <価格>980</価格>
</商品>
</発注伝票>
```

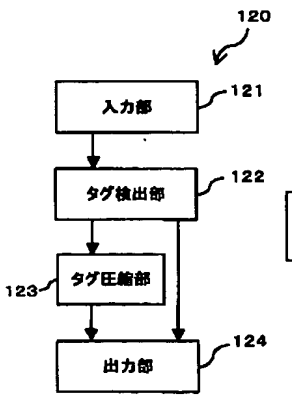
(B)

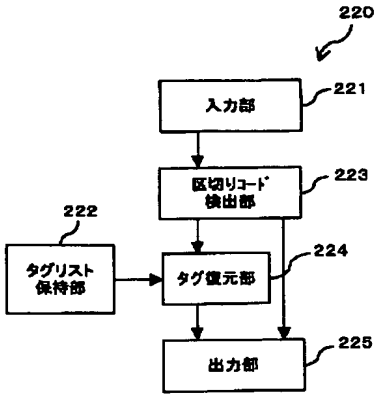
```
<発注伝票>
<発注者><名前></名前>
        <電話番号></電話番号>
</発注者>
<商品><メーカ></メーカ>
        <製品番号></製品番号>
        <製品名></製品名>
        <価格></価格>
</商品>
</発注伝票>
```

(C)

```
:
:, STUV,
  , 1111,
:
:, A社,
  , 1234,
  , ABCD,
  , 980,
:
:
```

【図４】





【図１８】

```
<商品 tag = 'タグリストA'>
, A社 , ,111A, ,980,
</商品>
```

識別子

【図６】

(A)

```
<商品><メーカ>A社</メーカ>
      <製品番号>1111</製品番号>
      <価格>980</価格>
</商品>
```

(B)

```
商品
メーカ
/メーカ
製品番号
/製品番号
価格
/価格
/商品
```

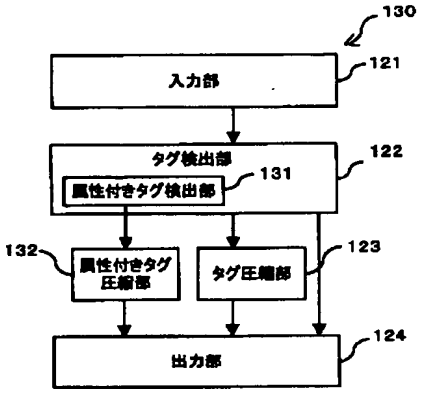
(C)

```
, , A社 , ,1111, ,980, ,
```

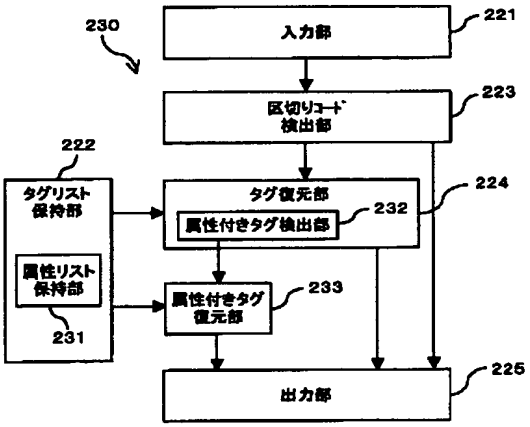
(D)

```
, , A社/,1111/,980//
```

【図７】



【図８】



【図９】

(A)

```
<商品>
  <メーカ>A社</メーカ>
  <製品番号 製品名='ABCD' 色='青'>1111</製品番号>
  <価格>980</価格>
</商品>
```

(B)

```
商品
メーカ
/メーカ
製品番号
=製品名
=色
/製品番号
価格
/価格
/商品
```

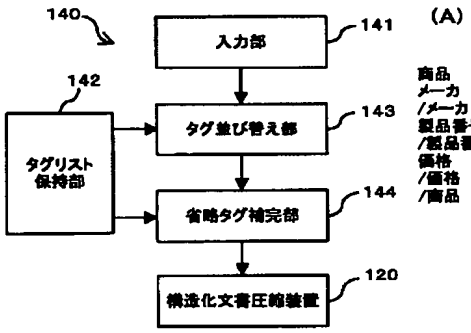
(C)

```
, , A社 , ,ABCD,青,1111, ,980, ,
```

(D)

```
, , A社 /,ABCD=青=1111/,980 //
```

【図１０】



【図１１】

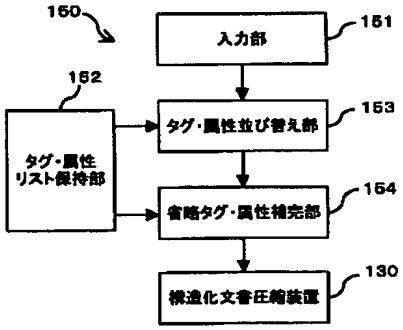
(B)

```
<商品><商品>300</価格>
      <製品番号>B7</製品番号>
</商品>
```

(C)

```
<商品><メーカ></メーカ>
      <製品番号>B7</製品番号>
      <価格>300</価格>
</商品>
```

【図１２】



【図１４】

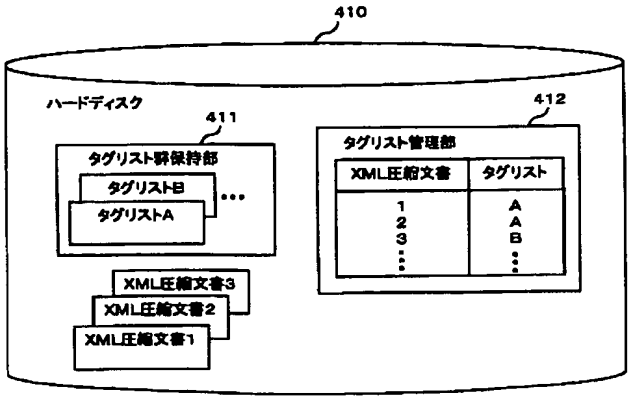
(A)

```
<発注伝票><発注者>佐藤</発注者>
  <商品><メーカ>A社</メーカ>
        <製品番号>111A</製品番号>   } サブ文書
        <価格>980</価格>
  </商品>
  <商品><メーカ>B社</メーカ>
        <製品番号>111B</製品番号>   } サブ文書
        <価格>500</価格>
  </商品>
  <商品><メーカ>C社</メーカ>
        <製品番号>111C</製品番号>   } サブ文書
        <価格>750</価格>
  </商品>
</発注伝票>
```

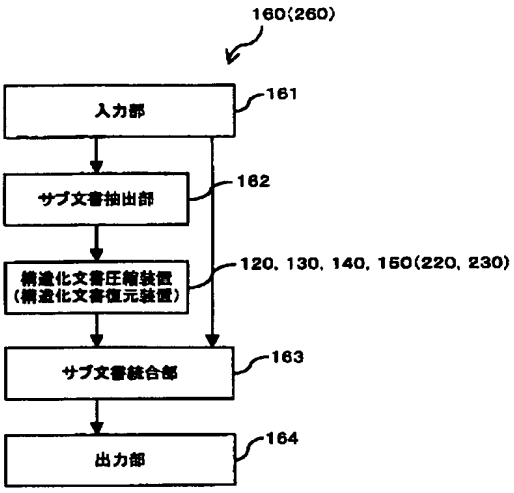
(B)

```
<発注伝票><発注者>佐藤</発注者>
  <商品>
    , A社 , ,111A, ,980,   } サブ文書
  </商品>
  <商品>
    , B社 , ,111B, ,500,   } サブ文書
  </商品>
  <商品>
    , C社 , ,111C, ,750,   } サブ文書
  </商品>
</発注伝票>
```

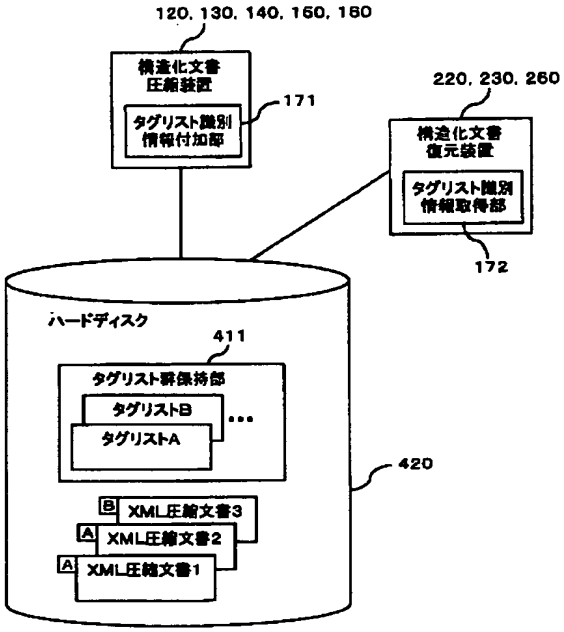
【図１５】



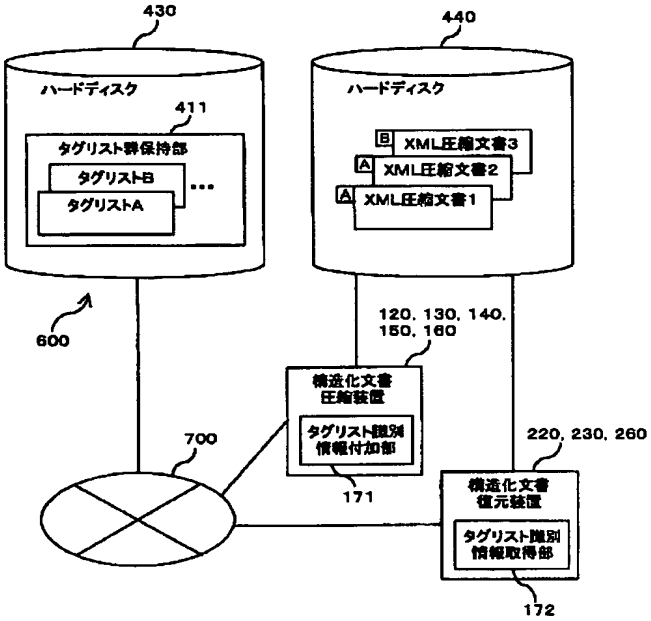
【図１３】



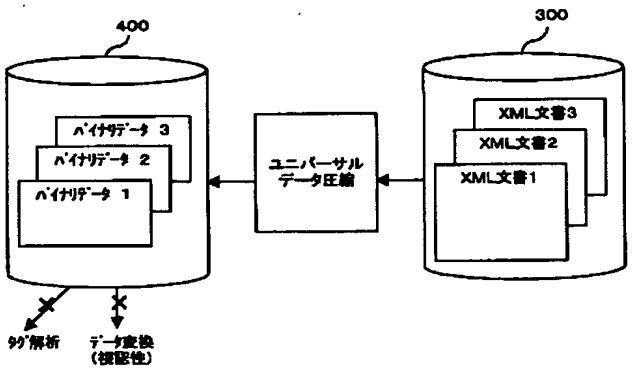
【図１６】



【図１７】



【図１９】



【図２０】